

UNITED STATES AIR FORCE  
SUMMER RESEARCH PROGRAM – 1995  
SUMMER FACULTY RESEARCH PROGRAM FINAL REPORTS

VOLUME 5B  
WRIGHT LABORATORY

RESEARCH & DEVELOPMENT LABORATORIES  
5800 Uplander Way  
Culver City, CA 90230-6608

Program Director, RDL  
Gary Moore

Program Manager, AFOSR  
Major David Hart

Program Manager, RDL  
Scott Licoscas

Program Administrator, RDL  
Gwendolyn Smith

Reproduced From  
Best Available Copy

Submitted to:

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH  
Bolling Air Force Base  
Washington, D.C.  
December 1995

19981218 059

REPORT DOCUMENTATION PAGE			Form Approved GSA FPMR (41 CFR) 101-11.6	
<p>Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing existing information, gathering material and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188).</p>				
1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE December, 1995	3. REPORT TYPE Final	AFRL-SR-BL-TR-98-0824	
4. TITLE AND SUBTITLE USAF Summer Research Program - 1995 Summer Faculty Research Program Final Reports, Volume 5B, Wright Laboratory			5. FUNDING NUMBERS	
6. AUTHORS Gary Moore				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Research and Development Labs, Culver City, CA			8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) AFOSR/NI 4040 Fairfax Dr, Suite 500 Arlington, VA 22203-1613			10. SPONSORING/MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES Contract Number: F49620-93-C-0063				
12a. DISTRIBUTION AVAILABILITY STATEMENT Approved for Public Release			12b. DISTRIBUTION CODE	
<p>13. ABSTRACT (Maximum 200 words)</p> <p>The United States Air Force Summer Faculty Research Program (USAF- SFRP) is designed to introduce university, college, and technical institute faculty members to Air Force research. This is accomplished by the faculty members being selected on a nationally advertised competitive basis during the summer intersession period to perform research at Air Force Research Laboratory Technical Directorates and Air Force Air Logistics Centers. Each participant provided a report of their research, and these reports are consolidated into this annual report.</p>				
14. SUBJECT TERMS AIR FORCE RESEARCH, AIR FORCE, ENGINEERING, LABORATORIES, REPORTS, SUMMER, UNIVERSITIES			15. NUMBER OF PAGES	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UL	

## **PREFACE**

Reports in this volume are numbered consecutively beginning with number 1. Each report is paginated with the report number followed by consecutive page numbers, e.g., 1-1, 1-2, 1-3; 2-1, 2-2, 2-3.

Due to its length, Volume 5 is bound in three parts, 5A, 5B and 5C. Volume 5A contains #1-23, Volume 5B contains reports #24-44 and 5C contains reports #45-64. The Table of Contents for Volume 5 is included in both parts.

This document is one of a set of 16 volumes describing the 1995 AFOSR Summer Research Program. The following volumes comprise the set:

### **VOLUME**

### **TITLE**

1.	Program Management Report
	<i>Summer Faculty Research Program (SFRP) Reports</i>
2A & 2B	Armstrong Laboratory
3A & 3B	Phillips Laboratory
4	Rome Laboratory
5A, 5B & 5C	Wright Laboratory
6A & 6B	Arnold Engineering Development Center, Wilford Hall Medical Center, and Air Logistics Centers
	<i>Graduate Student Research Program (GSRP) Reports</i>
7A & 7B	Armstrong Laboratory
8	Phillips Laboratory
9	Rome Laboratory
10A & 10B	Wright Laboratory
11	Arnold Engineering Development Center, Wilford Hall Medical Center and Air Logistics Centers
	<i>High School Apprenticeship Program (HSAP) Reports</i>
12A & 12B	Armstrong Laboratory
13	Phillips Laboratory
14	Rome Laboratory
15A&15B	Wright Laboratory
16	Arnold Engineering Development Center

## **SFRP FINAL REPORT TABLE OF CONTENTS**

**i-xiv**

<b>1. INTRODUCTION</b>	<b>1</b>
<b>2. PARTICIPATION IN THE SUMMER RESEARCH PROGRAM</b>	<b>2</b>
<b>3. RECRUITING AND SELECTION</b>	<b>3</b>
<b>4. SITE VISITS</b>	<b>4</b>
<b>5. HBCU/MI PARTICIPATION</b>	<b>4</b>
<b>6. SRP FUNDING SOURCES</b>	<b>5</b>
<b>7. COMPENSATION FOR PARTICIPATIONS</b>	<b>5</b>
<b>8. CONTENTS OF THE 1995 REPORT</b>	<b>6</b>

### **APPENDICIES:**

<b>A. PROGRAM STATISTICAL SUMMARY</b>	<b>A-1</b>
<b>B. SRP EVALUATION RESPONSES</b>	<b>B-1</b>

### **SFRP FINAL REPORTS**



**GAS CHROMATOGRAPHY/MASS SPECTROMETRY OF  
PERFLUOROPOLYALKYLETHER BASED LUBRICATING OILS**

**David W. Johnson  
Associate Professor  
Department of Chemistry**

**University of Dayton  
300 College Park  
Dayton, OH 45469-2357**

**Final Report for:  
Summer Faculty Research Program  
Wright Laboratory**

**Sponsored by:  
Air Force Office of Scientific Research  
Bolling Air Force Base, Washington, D.C.**

**and**

**Wright Laboratory**

**August, 1995**

**GAS CHROMATOGRAPHY/MASS SPECTROMETRY OF  
PERFLUOROPOLYALKYLETHER BASED LUBRICATING OILS**

**David W. Johnson  
Associate Professor  
Department of Chemistry  
University of Dayton**

**Abstract**

Gas chromatography/mass spectrometry has been used to separate and identify the individual oligomers in the perfluoropolyalkylether based lubricants; Krytox 143 AZ and two low molecular weight fractions of Krytox 143 AC. Based on the mass spectrometry, the molecular weight of the individual oligomers can be determined. Gas chromatography/mass spectrometry has also been used to determine the degradation product of a lubricant additive in another perfluoropolyalkylether based lubricant; Fomblin Z. Mass spectrometric data indicates that the degradation pathway proceeds through the formation of an acid fluoride as the first step. The second step in the degradation adds a ketone functional group to the molecule.

## Introduction

The Air Force, as one of its long term goals is looking to develop a new generation of jet engines. These engines will have double the thrust to weight ratio of the current generation of jet engines. In order to achieve greater thrust, these engines will operate at much higher temperatures than current engines. Operation of these engines at higher temperatures requires new high temperature materials for many critical components.

One of the major challenges posed by a higher temperature is in the area of lubrication. The many moving parts will require lubrication. Lubricants must be able to stand up to the substantially high temperatures than is possible using currently available lubricants. These fluids will be required to be liquids, chemically inert in contact with air and metals, and act as lubricants over temperatures from -40 to 700°F. The design of lubricants for use over the wide range of temperatures poses a challenge to scientists in the study of materials.

Conventional liquid lubricants are based on a hydrocarbon base stock which degrades readily at high temperatures. Lubricants for advanced jet engines will likely be based on a fluorinated base stocks. Among the materials currently under consideration are series of perfluoropolyalkylethers (PFPAEs) which have a wide liquid range and are stable in the presence of oxygen to temperatures substantially higher than conventional lubricants<sup>1</sup>

In this report, the application of gas chromatography/mass spectrometry to the analysis of perfluoropolyalkylethers will be

discussed. The molecular weights of the various oligomers found in Krytox 143 AZ and two fractions of supercritical fluid extracted Krytox 143AC have been determined. Gas chromatography/mass spectrometry has also been applied to the identification of degradation products of an additive in the perfluoroalkylether; Fomblin Z.

## Experimental Section

### Gas Chromatography/Mass Spectrometry (GC/MS)

GC/MS data was obtained using a Kratos 1-H double focussing mass spectrometer coupled with a Hewlett-Packard 5890 gas chromatograph. Chromatographic conditions were optimized for each set of samples and will be specified along with the results. Mass spectra were collected as raw data, in a scanning mode. Samples were ionized using electron impact with 40 EV electrons and the ions formed were accelerated using an 8kV potential. The data was processed with the Kratos Mach 3 software package using a 13 point Savitski-Golay smoothing algorithm. The instrument was calibrated using perfluorononyltriazine for masses to 1485 and a mixture of perfluorononyltriazine and UltraMark 1621 for samples with masses to 2100AMU.

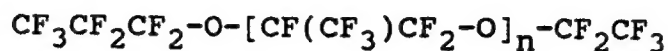
### Materials

Krytox 143 AZ samples, the two fractionated samples of Krytox 143 AC and the Fomblin Z samples with added additive were supplied by WL/MLBT for this study. Samples were dissolved in Freon 112 as a solvent before analysis. Samples were made up to be ~1% sample in the solvent.

## Results

### Gas Chromatography/Mass Spectrometry of Krytox 143AZ

Krytox 143 is a perfluoropolyalkylether prepared by the polymerization of perfluoropropylene oxide. Under the conditions of its preparation, the structure of the molecule is shown below.



where n varies substantially. Of particular interest is Krytox 143AC which has an average molecular weight of ~7,000 and a very broad distribution of molecular weights. Gas chromatography/mass spectrometry data for Krytox samples was obtained in the hope of determining details of the molecular weight distribution. The samples investigated include Krytox 143 AZ, the lowest molecular weight Krytox available and the two lowest molecular weight fractions of Krytox 143 AC which was extracted using supercritical fluids by Phasex for WL/MLBT. These samples were chosen because they contained only relatively low molecular weight materials which would be amenable to separation by GC.

The GC/MS of krytox 143AZ was measured using an initial temperature of 40°C for two minutes, then increasing the temperature at a rate of 8°C/min to a final temperature of 300°C. These GC conditions allowed the peaks from the low boiling components to be observed. The conditions did result in some broadening of the higher molecular weight peaks. Under these conditions, a solvent delay of 1.75 minutes was required in order to allow the solvent to elute from the column without burning the filament.

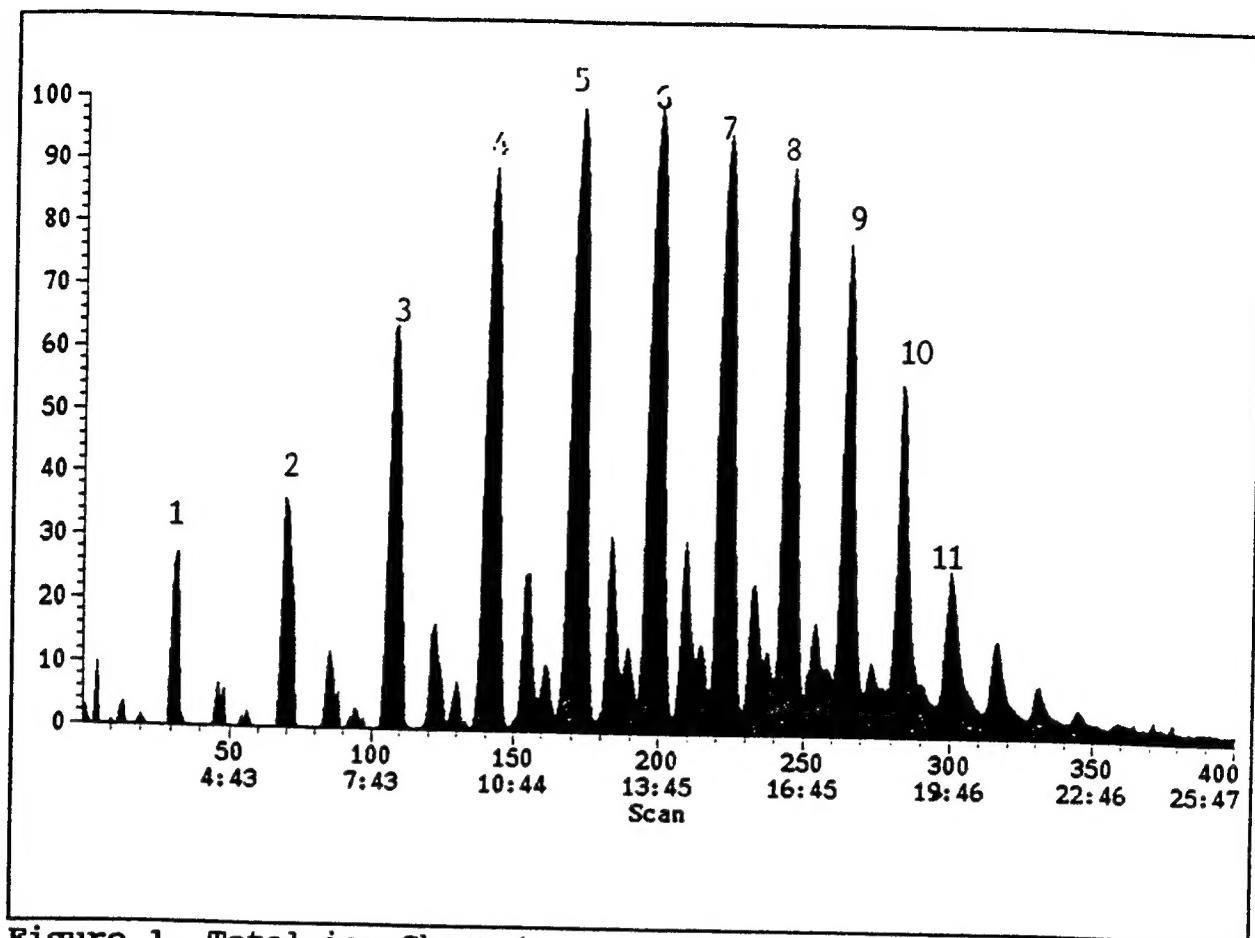


Figure 1. Total ion Chromatogram of Krytox 143AZ.

The GC/MS total ion chromatogram for Krytox 143 AZ is shown in Figure 1. The chromatogram consists of a series of triads arrange in a bell shaped curve. The earliest of the peaks are well resolved from their neighboring peaks with the resolution decaying for later peaks in the chromatogram. The distribution of the peaks is in the bell shape, often seen for polymeric materials<sup>2</sup>. While a complete quantitative analysis of this chromatogram is not possible without samples of the individual components, the peak areas will approximate the mass percentages due to the similarity of the various components.

The mass spectra of the major peaks have the same general

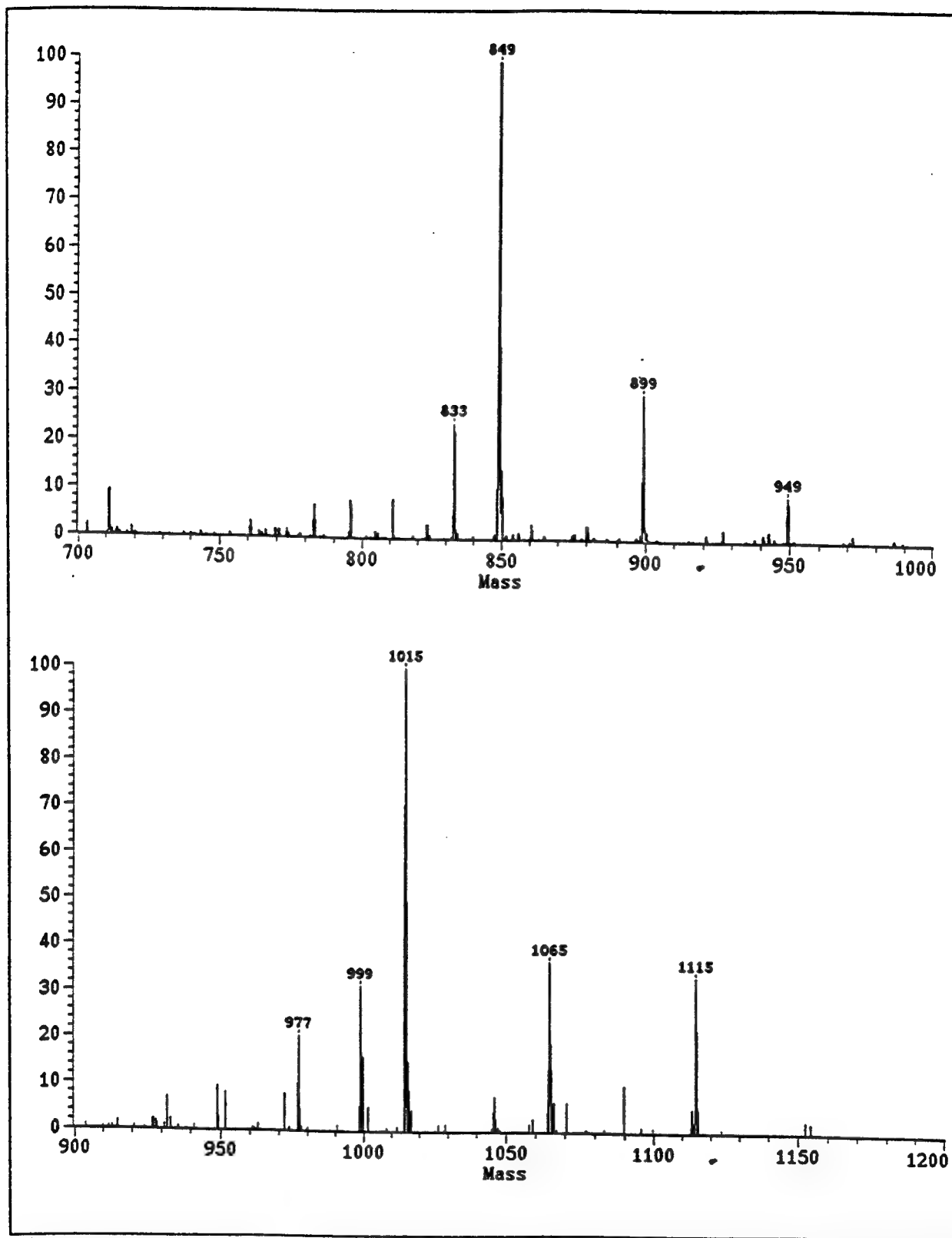


Figure 2. The high mass portion of the mass spectra of the first two major peaks in the GC/MS total ion chromatogram of Krytox 143AZ.

Table I. Mass Spectrometry Data for the Dominant Peaks in The GC/MS Analysis of Krytox 143AZ.

Peak Number <sup>a</sup>	Parent Ion (AMU) <sup>b</sup>	M-F observed (calc) <sup>c</sup>	M-CF <sub>3</sub> observed (calc)	M-C <sub>2</sub> F <sub>5</sub> observed (calc)	M-C <sub>2</sub> F <sub>5</sub> O observed (calc)
1	968	949 (949)	899 (899)	849 (849)	833 (833)
2	1134	1115 (1115)	1065 (1065)	1015 (1015)	999 (999)
3	1300	1281 (1281)	1231 (1231)	1181 (1181)	1165 (1165)
4	1466	1447 (1447)	1397 (1397)	1347 (1347)	1331 (1331)
5	1632	1612 (1613)	1563 (1563)	1513 (1513)	1497 (1497)
6	1798	1778 (1779)	1728 (1729)	1678 (1679)	1662 (1663)
7	1964	1943 (1945)	1893 (1895)	1843 (1845)	1827 (1829)
8	2130	2111 (2111)	2061 (2061)	2011 (2011)	1997 (1997)

<sup>a</sup> Peak number as identified in Figure 1.

<sup>b</sup> Calculated, we have not been able to identify the parent ion in any of the GC/MS experiments.

<sup>c</sup> The deviation between the observed mass and calculated mass can be explained by considering that the instrument was calibrated to a mass of 1485 AMU.

features, including very intense peaks at low mass, indicating extensive fragmentation. Examination of the high mass region shows the absence of a parent ion peak. Parent ion peaks are normally not observed for perfluorinated aliphatic species. The mass spectra of the first two major peaks (scans 29-32 and 68-71 are shown in Figure 2). In Figure 2a, the highest observed masses are 949, 899, 849 and 833 AMU. These peaks correspond to the loss of F, CF<sub>3</sub>, C<sub>2</sub>F<sub>5</sub>



and  $C_2F_5O$ , respectively from a parent mass of 968AMU. The parent ion is not observed for these molecules, but the presence of a peak due to M-F is similar to the pattern observed for many other highly fluorinated molecules. Figure 2b shows highest observed masses of 1115, 1065, 1015 and 999AMU, corresponding to the loss of the same fragments from a parent mass of 1134 AMU. At lower masses, ions formed by the fragmentation of the perfluoropolyalkylether on either side of the ether oxygen can be observed. There is also evidence that while fragmentation from either end of the molecule is possible, fragmentation from the  $C_2F_5$  end of the molecule is favored. Mass spectra for the other members of the dominant series are summarized in Table 1. The mass spectra show the expected series of mass peaks which differ from neighboring chromatographic peaks by 166 mass units.

Identification of the two series of minor components seen in the GC/MS results for Krytox 143AZ is more difficult due to the small concentration of the compounds and the extensive fragmentation seen in this type of compound. The first of these components are thought to result from an alternate chain termination step, known to occur in the synthesis of Krytox. The final series of components arises due to the presence of a small amount of residual hydrogen in all of the Krytox 143 materials. Confirmation of the hydrogen containing can be seen in the mass spectra (Figure 3) of these peaks. The presence of a large peak at 101 AMU is indicative of the presence of the  $CF_3CHF$  group which has been shown to contain the hydrogen in Krytox.

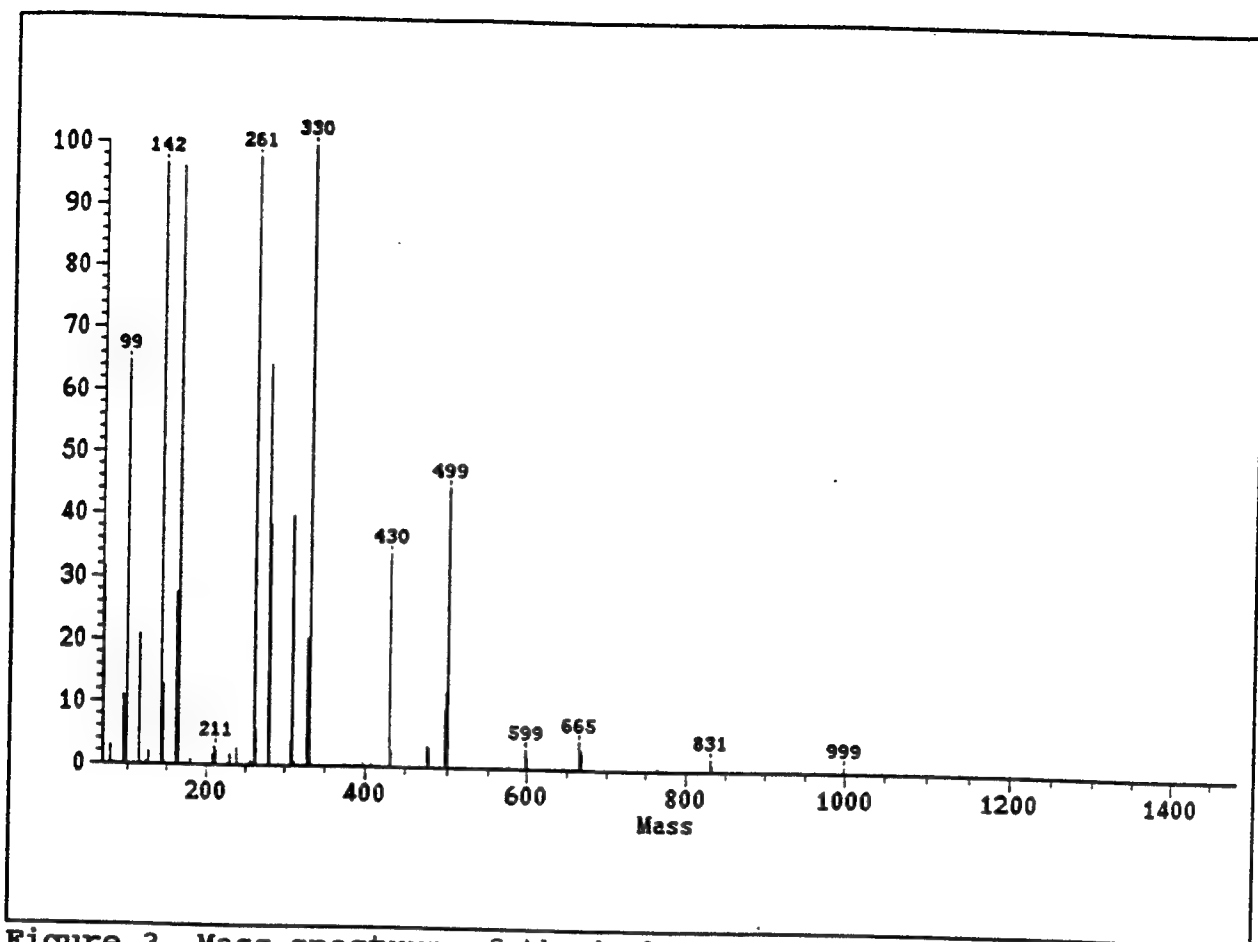


Figure 3. Mass spectrum of the hydrogen containing peak of Krytox 143 AZ.

Further confirmation of the identification of this series of peaks is from the GC/MS of a sample of Freon E-6. Freon E-6 is a perfluoropropylene oxide polymer which terminates in the  $\text{CF}_3\text{CFH}$  unit as is postulated for the hydrogen containing species in Krytox 143AZ. The GC/MS total ion chromatogram for this sample is shown in Figure 4. In addition to the main peak due to Freon E-6, we see minor peaks due to other higher and lower molecular weight members of the Freon E series. These features are at the same retention time that are observed for the third series in Krytox 143 AZ. If we compare the mass spectra of the main peak in Freon E-6 with the mass spectrum of third series of components (Figure 3 Figure 5a) we

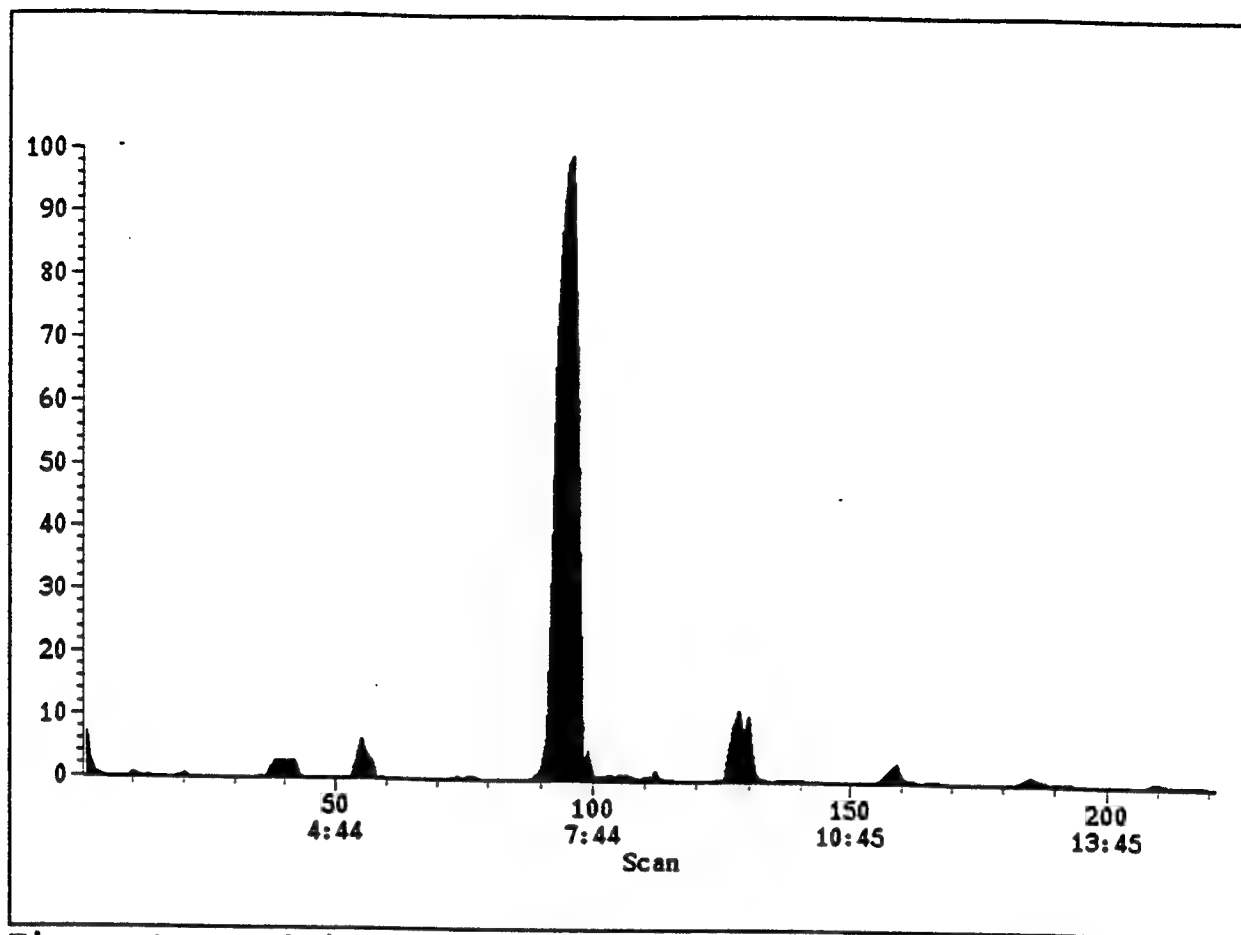


Figure 4. Total ion chromatogram for a sample of Freon E-6.

see definite similarities. If we also examine the high mass region of the mass spectrum of Freon E-6 (Figure 5b) we see that the highest mass peak observed is mass 1097 which corresponds to the loss of fluorine from the parent mass of 1116. Other features include masses of 947 (loss of  $\text{CF}_3$ ) and 999 (loss of  $\text{CF}_3\text{CFHO}$ ) from the parent ion. These mass spectra also confirm our assignment of the mass spectrum of the dominant series in Krytox 143AZ by indicating that the M-F peak is the highest mass peak observable in compounds of this type. The retention time of the main peak is between the second and third peaks of the main series of Krytox 143AZ (mass 1134 and 1300), suggesting that the presence of single

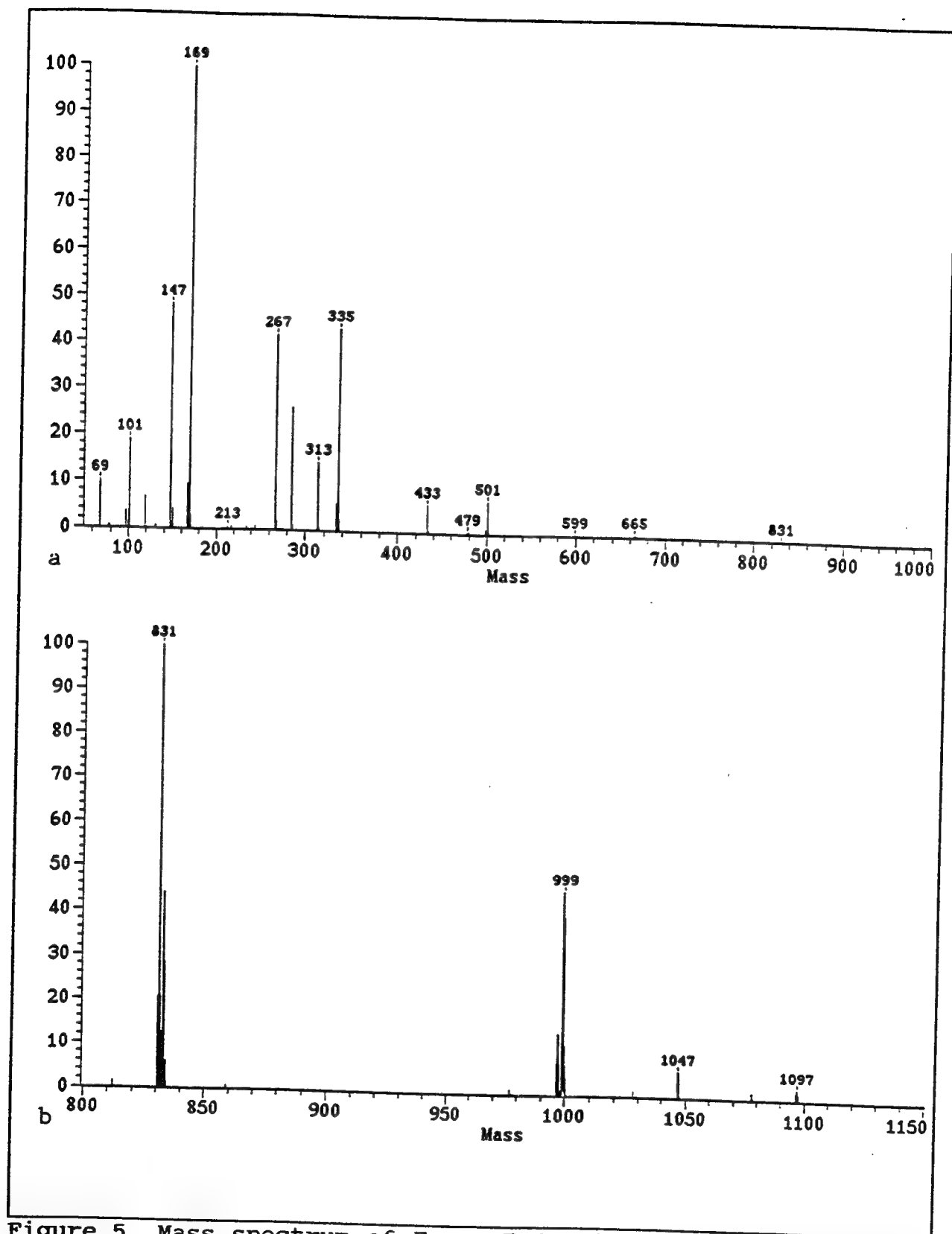


Figure 5. Mass spectrum of Freon E-6. a) the major features of the mass spectrum; b) the high mass region of the mass spectrum.

hydrogen in the molecule increases the molecules interaction with the stationary phase slightly.

#### Gas Chromatography/Mass Spectrometry of Krytox 143AC

Krytox 143 AC is similar in structure to Krytox 143 AZ, however the average molecular weight is substantially higher. Previous work has shown that partial separation of Krytox 143 AC by GC can be obtained, but the higher molecular weight components do not elute from the column. To begin to determine the molecular weight distribution in Krytox 143-AC, we have examined the GC/MS of the two lowest molecular weight fractions of Krytox 143 AC which has been supercritical fluid extracted by Phasex. The GC/MS total ion chromatogram of these two samples are shown in Figure 6.

The total ion chromatogram of the lowest molecular weight fraction of Krytox 143 AC is very similar to that of Krytox 143 AZ but contains some material of higher molecular weight. The mass spectra of the first several peaks are identical with those of Krytox 143AZ and so a detailed interpretation of these spectra will not be presented. The other difference is the smaller area of the hydrogen containing species in these samples consistent with the smaller quantity of hydrogen present within Krytox 143AC.

The total ion chromatogram for the second fraction of Krytox 143 AC is of considerably higher molecular weight than the first fraction. There is substantial overlap in the high molecular weight portion of the total ion chromatogram. Very small amounts of the low molecular weight components are observed at retention times consistent with the first fraction. However a substantial fraction

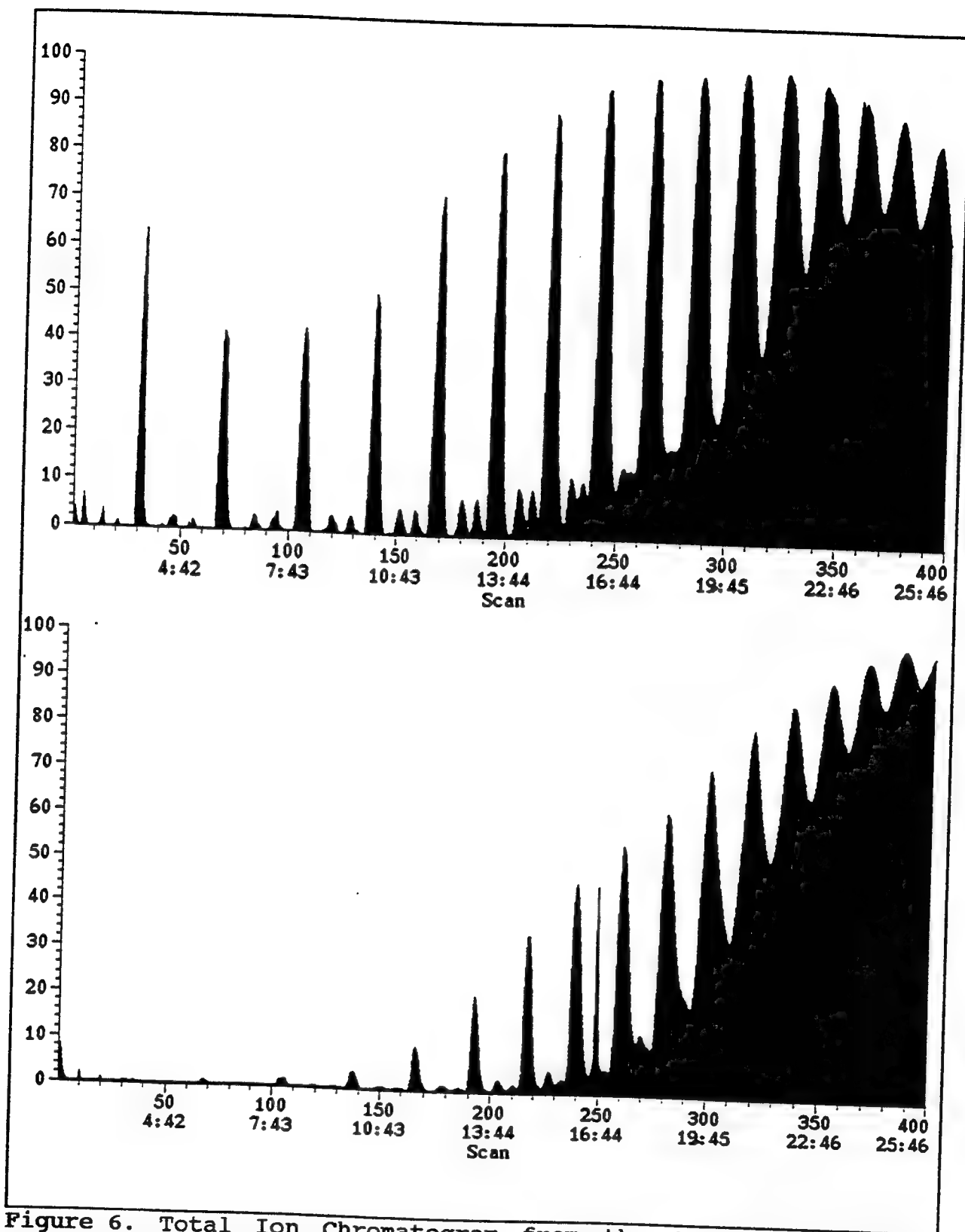


Figure 6. Total Ion Chromatogram from the GC/MS Analysis of Supercritical Fluid Extracted Krytox 143-AC; a) Sample 7-1, the lowest M.W. sample and b) Sample 7-2 the second fraction.

Table II. Comparison of Retention Times for Krytox 143AZ and two fractions of Krytox 143AC.

Peak Number	Molecular Weight	Scan Numbers Krytox 143 AZ	Scan Numbers Krytox 143 AC-1	Scan Numbers Krytox 143 AC-2
1	968 <sup>a</sup>	30-32	30-32	N.O.
2	1134 <sup>a</sup>	68-71	68-71	68-69
3	1300 <sup>a</sup>	106-109	106-109	106-108
4	1466 <sup>a</sup>	139-143	138-143	138-140
5	1632 <sup>a</sup>	169-174	168-173	168-170
6	1798 <sup>a</sup>	194-201	193-200	192-197
7	1964 <sup>a</sup>	220-226	219-225	218-222
8	2130 <sup>a</sup>	242-248	241-249	239-245
9	2296 <sup>b</sup>	263-268	260-268	259-266
10	2462 <sup>b</sup>	282-286	281-292	280-288
11	2628 <sup>b</sup>	300-305	300-315	299-308
12	2794 <sup>b</sup>	315-319	318-330	309-328
13	2960 <sup>b</sup>	330-332	330-351	329-351
14	3126 <sup>b</sup>	N.O.	351-365	351-366
15	3292 <sup>b</sup>	N.O.	365-384	367-390
16	3458 <sup>b</sup>	N.O.	385-400	391-405

<sup>a</sup> Based on the observed masses of the M-F, M-CF<sub>3</sub>, M-C<sub>2</sub>F<sub>5</sub> and M-C<sub>2</sub>F<sub>5</sub>O ions

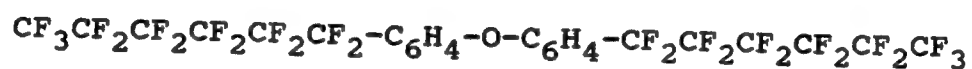
<sup>b</sup> Based on the extrapolation of the series by the addition of successive monomer units(mass 166).

of this sample does not elute from the column as resolvable peaks, but continues to elute as a broad unresolved mound. Due to the overlap of this chromatogram with the low molecular weight sample and Krytox 143 AZ, however we can assign masses to the oligomers represented by the chromatographic peaks by extending the series determined from Krytox 143 AZ. This data is presented in Table 2.

The retention times for the various peaks are sufficiently close that identification of the molecular weight of the particular species is simplified. Based on this data, and comparison of the GC total ion chromatograms with supercritical fluid chromatograms (which separated all of the different components), masses can be assigned to all of the different oligomers of Krytox 143 AC.

#### Gas Chromatography/Mass Spectrometry Studies of Additive Degradation

The gas chromatography/mass spectrometry of an additive in Fomblin Z, stressed under various conditions was studied in order to determine the products of degradation. Fomblin Z is a The additive studied was a diphenylether with fluorinated chains attached in order increase the solubility of the additive in perfluoropolyalkyl ethers. The structure of this additive is shown below.



The additive was stressed in the presence of metals and oxygen at temperatures of 329°C, 343°C and 369°C. The total ion chromatograms for samples stressed at these three temperatures are show the degradation of the additive in two steps. The chromatographic peaks are sharp and well resolved, indicating the formation of single specific compounds in the degradation scheme. The sample stressed at 329°C shows only the additive in addition to the broad background of Fomblin Z. At 343°C the additive and a first decomposition product can be seen; while at 369°C the additive and two decomposition products are observed.



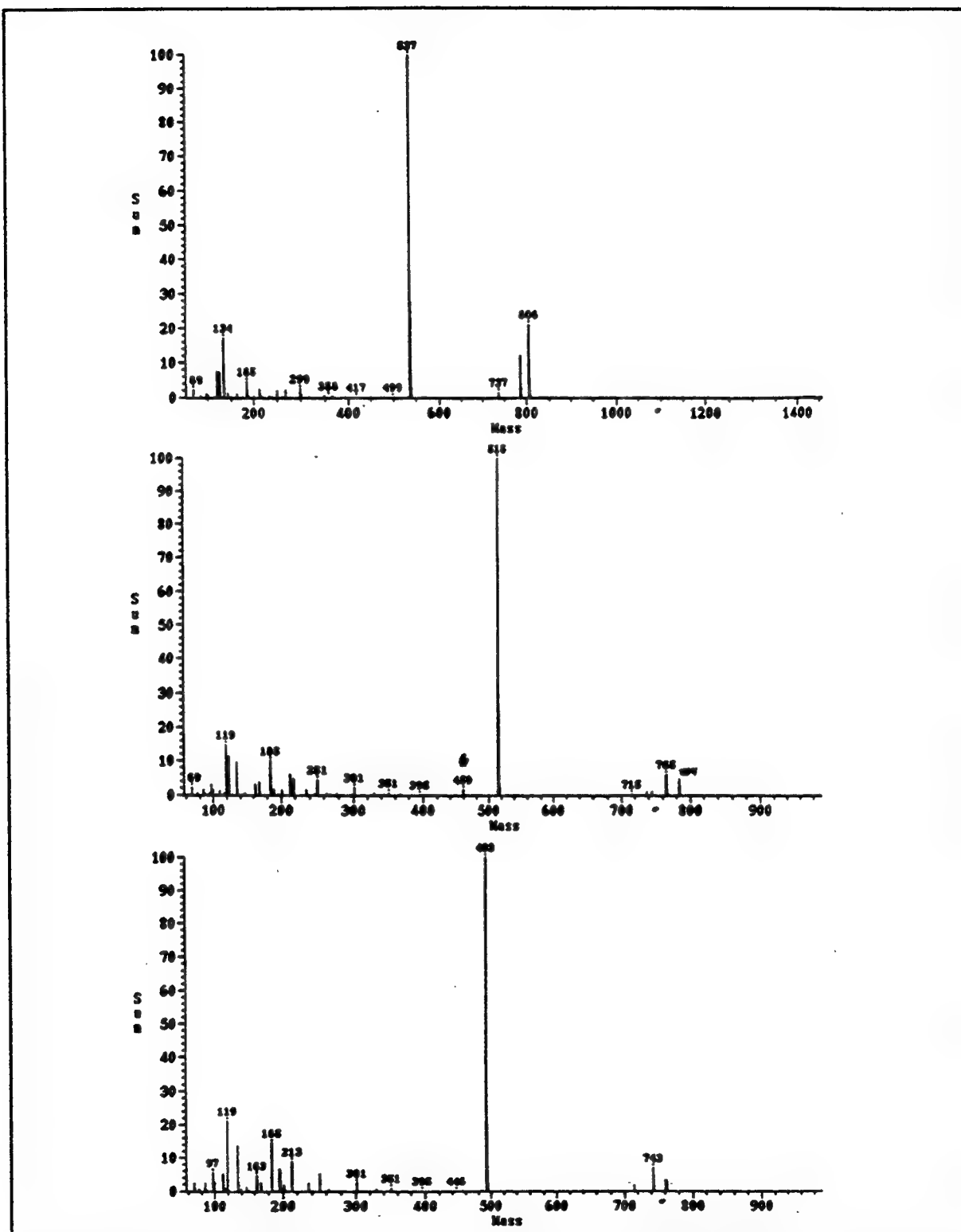


Figure 7. Mass Spectra of samples of a diphenylether additive in Fomblin Z; a) stressed at 329°C, b) stressed at 343°C, c) stressed at 369°C

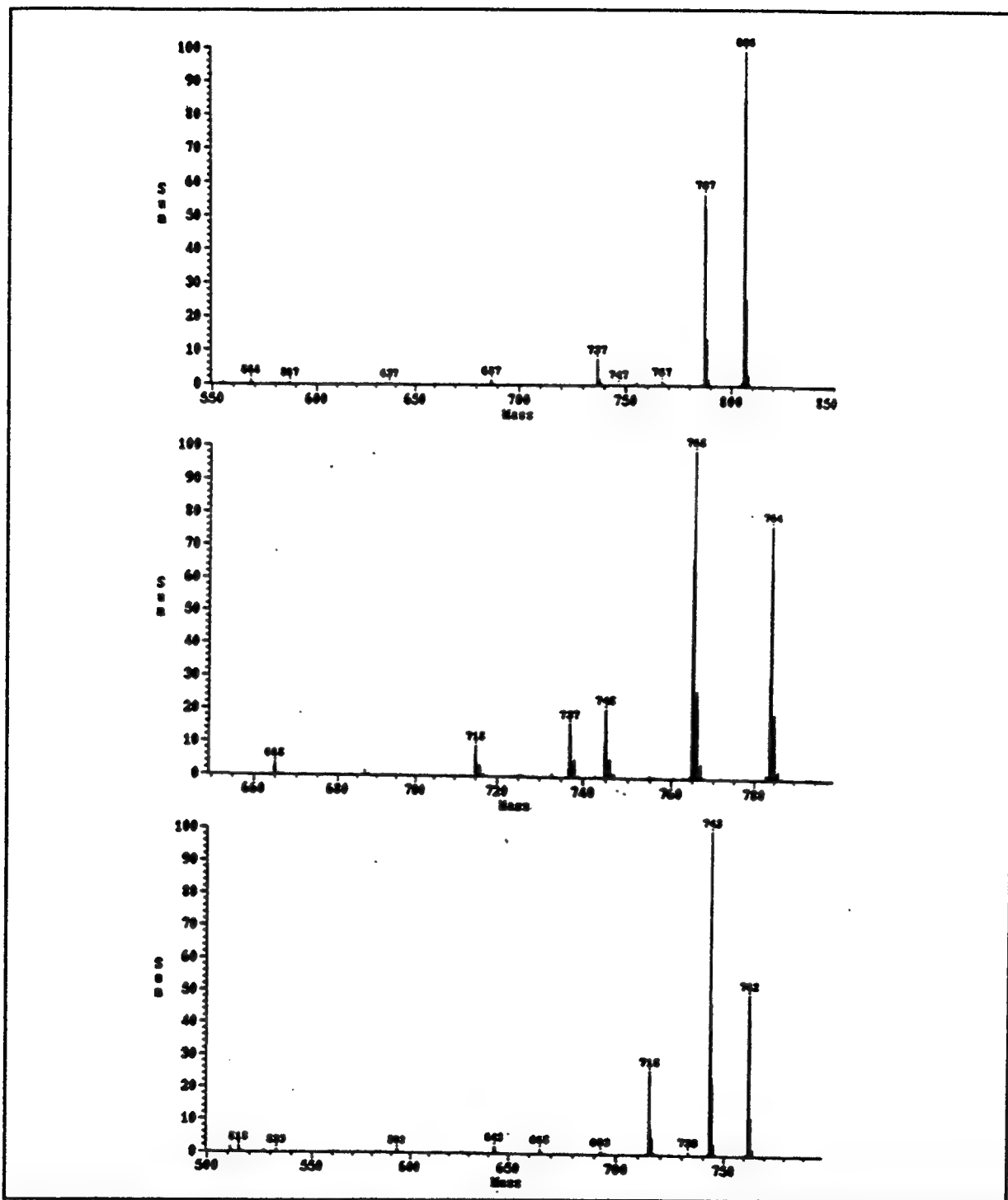


Figure 8. High Mass Portion of the Mass spectra of the Additive and Additive Decomposition Products in Fomblin Z; a) mass spectrum of the additive, b) mass spectrum of the first decomposition product and c) mass spectrum of the second decomposition product.

The mass spectra of the additive and the two decomposition products are shown in Figure 7. The details of the mass spectra in the vicinity of the parent ion are shown in Figure 8. If the mass spectrum of the additive, stressed at 329°C is examined (see Figure 7a, 8a), the parent ion is observed at 806 mass units. The dominant fragments in the mass spectrum are found at 787, 737 and 537 mass units corresponding to the loss of F, CF<sub>3</sub> and C<sub>5</sub>F<sub>11</sub>. This pattern is similar to the pattern observed for other perfluorinated aromatic compounds where the dominant ion results from cleavage of the molecule leaving a CF<sub>2</sub> attached to the aromatic ring.

If we examine the mass spectrum of the first of the decomposition products we see that the parent ion is observed at a mass of 784 mass units. The mass difference between this molecule and the original additive suggests the replacement of two fluorine atoms with an oxygen atom. A pattern of fragmentation similar to the original additive is observed for this compound showing ions due to the loss of F, CF<sub>3</sub> and C<sub>5</sub>F<sub>11</sub> from the parent. There is an additional peak observed at M-47 which corresponds to the loss of COF from the molecule. This observation requires the formation of an acyl fluoride as the decomposition product.

Analysis of the mass spectrum of the second decomposition product again shows the loss of 22 mass units from the first decomposition product. This observation suggests that two more fluorine atoms are replaced with an oxygen. The assignment of a structure to this molecule is substantially more difficult. The presence of fragments due to the loss of F, CF<sub>3</sub> and C<sub>5</sub>F<sub>11</sub> requires

that the reaction has occurred on the same perfluorinated chain as the first reaction. Detailed examination of the fragments formed show losses of 97, 147 and 247 mass units corresponding to the loss of  $\text{CF}_2\text{COF}$ ,  $\text{C}_2\text{F}_4\text{COF}$  and  $\text{C}_4\text{F}_8\text{COF}$  from the parent ion respectively. This set of observations suggests that the second oxidation site is adjacent to the aromatic ring on the same half of the molecule.

### Conclusions

Gas chromatography/mass spectrometry has been used to separate and identify many of the individual components of Krytox 143AZ and Krytox 143AC. The mass spectra data will allow the assignment of actual molecular weights the various components and allow future studies of the degradation mechanisms for this lubricant.

The analysis of additives in stressed Fomblin Z samples has illustrated the power of gas chromatography/mass spectrometry in determining degradation mechanisms. The additive has been shown to decompose in specific manners to give easily characterized compounds.

### References Cited

1. Jones, W.R. Jr.; Paciorek, K.J.L.; Ito, T.I. and Kratzer, R.H. Ind. Eng. Chem. Prod. Res. Dev.; 22, 1983, 166-70; Snyder, C.E. Jr.; Gschwender, L.J. and Tamborski, C. Lubr. Eng.; 37, 1981, 344-9.
2. J. F. Sullivan in "Modern Practice of Gas Chromatography, Second Edition"; R.L. Grob, Editor; John Wiley and Sons, 1985, 721-757.

**CONTAINERLESS PROCESSING  
OF  
SINGLE (PST) CRYSTALS OF LAMELLAR TiAl**

**Bimal K. Kad  
Research Scientist  
Department of Applied Mechanics & Engineering Sciences**

**University of California-SanDiego  
Mail Code-0411, 9500 Gilman Drive  
La Jolla, CA 92093-0411**

**Final Report for:  
Summer Faculty Research Program  
Wright Patterson Air Force Laboratory**

**Sponsored by:  
Air Force Office of Scientific Research  
Bolling Air Force Base, DC**

**and**

**Wright-Patterson Laboratory**

**September 1995**

# CONTAINERLESS PROCESSING OF SINGLE (PST) CRYSTALS OF LAMELLAR TiAl

Bimal K. Kad  
Research Scientist  
Department of Applied Mechanics & Engineering Sciences  
University of California-San Diego

## Abstract

Containerless growth of single poly-synthetically twinned (PST) crystals of  $\alpha_2+\gamma$  lamellar crystals was carried out with a view to obtaining oriented single crystals of Ti-49at%Al composition, that have their lamellar plates aligned parallel to the growth direction. The anisotropy of crystal growth, in the precursor  $\alpha$ -hcp phase, was employed to orient the basal plane parallel to the growth direction, thereby yielding the desired lamellar orientation. Experimentally, the liquidus composition was zone leveled in the initial transients, to ensure  $L \rightarrow \alpha$  solidification while avoiding  $L \rightarrow \beta$  solidification.

# CONTAINERLESS PROCESSING OF SINGLE (PST) CRYSTALS OF LAMELLAR TiAl.

Bimal K. Kad

## § 1. Introduction

Understanding the physical phenomena at the single crystal level is an essential prerequisite to the understanding the polycrystal aggregate, with all the inherent, and additional complexities of scale and grain boundary effects. This initial objective of characterizing phenomenon at the single crystal level, must begin with sophisticated experimental setups that can produce reasonable volumes (of the order of 10mm diameter x 100mm length) of single crystal of any composition and melting point. These goals are met by crucible-based Czochralski crystal growth techniques, for unreactive simple metals and alloys, and are currently of commercial significance. The use of silicon single crystals in the semiconductor industry is one example of this mature technology.

Current development work for advanced high temperature structural applications is focused on reactive, refractory alloys of the aluminide family. The specific materials of interest, in the current context, are Ti-Al based alloys, currently being targeted for the hot section turbine blades of jet engines. Since these alloys are thermodynamically quite reactive in the liquid state, their processing into useful shapes for materials research development, let alone useful engineering shapes, is complicated. Powder metallurgy, as is frequently employed, has the inherent problem of working with enormous reactive surface areas, at sintering temperatures. Conventional solidification through the liquid state, is bound to pick up contaminants due to molten metal contact with ceramic crucibles. Most ceramic crucibles begin to degrade above 1600°C, especially so when in contact with metals as Al, Ti, Zr, Cr, Mn etc.

Containerless induction melting, where the molten zone is supported by solid of like composition, is one suitable process to overcome these limitations. However, due to the non-contact requirement, the inductively coupled turbulent melt is difficult to monitor and control, such that steady state mass-transport through the molten zone may only be obtained by human intervention and endless trial and error experimentation. More recently image based computer control, of the liquid zone shape, is employed to eliminate trial and error practices [1,2]

### § 1.1 statement of experimental objectives

The material of interest is two phase  $\alpha_2$ -Ti<sub>3</sub>Al +  $\gamma$ -TiAl, with a lamellar microstructure. Lamellar titanium-rich TiAl initially solidifies as disordered  $\alpha$ -hcp, see phase diagram in figure 1,

and undergoes solid-state transformations  $\alpha \rightarrow \gamma$  and  $\alpha \rightarrow \alpha_2 + \gamma$  to produce fine scale (1-2 $\mu\text{m}$  wide) lamellar microstructure, where each grain is a flat slab bounded by large close packed  $\{111\}\gamma \parallel (0001)\alpha_2$  planes. Lamellar material has more intrinsic resistance to deformation across laminate boundaries (hard mode) than to deformation which is parallel to the boundaries (soft mode). This fact, combined with the shapes of the lamellae (small slip distances across the lamellae and large slip distances parallel to the lamellae) causes a marked, orientation dependent, plastic anisotropy in lamellar material. Of particular interest is the mechanical response of this alloy microstructure when lamellar plates are aligned parallel to the tensile axis. This orientation offers the best combination of strength and ductility, and current experimental efforts are intended to obtain this orientation (or texture) in both PST-single crystal and the polycrystalline materials.

Current objective is to define the processing envelope for producing single crystals routinely. Towards this end, current practices are examined critically, and a systematic study, of the critical variables, was initiated to refine existing processing techniques. Furthermore, the inherent mechanical and structural anisotropies of lamellar TiAl dictate the need for orientation control or 'seeding' options during crystal growth, to facilitate the characterization of orientation specific properties in slender crystal cross-sections. Suitable means are employed to introduce texture (i.e., basal plane parallel to the growth direction) in the  $\alpha$ -hcp precursor phase

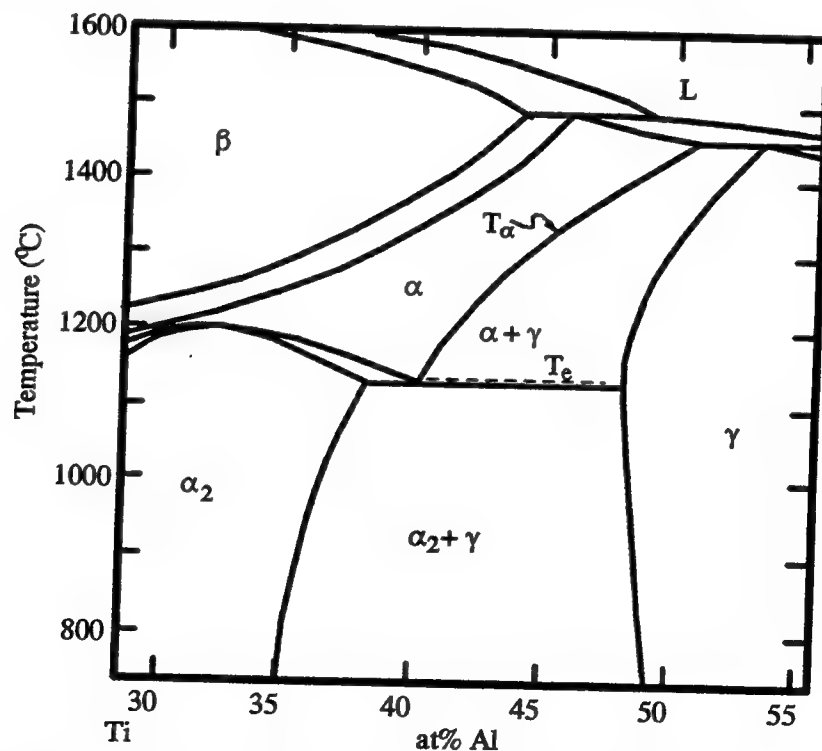


Figure 1. Experimental Ti-Al phase diagram of interest. For the Ti-(46-50at%)Al compositions of interest, the initial liquidus solidifies as either  $\beta$ , or  $\alpha$ -hcp, which undergo solid state transformations to yield the  $\alpha_2$ -Ti<sub>3</sub>Al +  $\gamma$ -TiAl laminate morphology.



## § 2. Details of the Experimental Setup

A system for single crystal growth employing containerless directional solidification has been developed. Induction heating is utilized to melt a liquid zone, the stability of the liquid zone shape being determined primarily by surface tension effects. Only a fraction of the charge material is melted at any one time, with mass transport through the shaped induction coil facilitated by continuous melting and freezing at the solid-liquid and liquid-solid interfaces respectively, figure 2. The ingot bar is positioned vertically through the stationary induction coil, and is typically thought of as two solid pieces with a molten zone in between. DC motors are used (separately or in synchronously) to rotate the melting (S-L) and freezing (L-S) interfaces, and to translate the charge assembly through the induction zone. The inherent requirement is to control the solidification parameters at the freezing (L-S) interface, in order to control the metallurgical structure of the growing crystal.

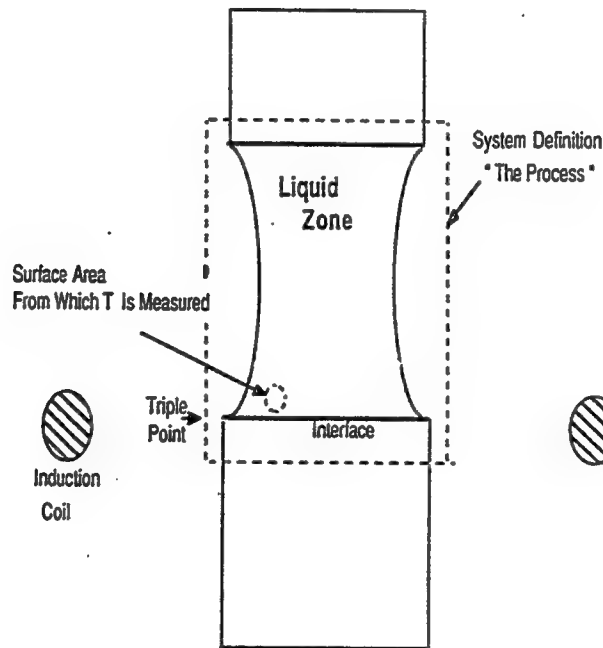


Figure 2. Schematic diagram of the containerless crystal growth process.

### § 2.1 phase-diagram considerations

The  $L+\alpha$  phase field is straddled between  $L+\beta = \alpha$ , and  $L+\alpha = \gamma$  peritectics at the Ti-rich and Al-rich composition ends, figure 1. This narrow composition window for  $L \rightarrow \alpha$  solidification, necessitates accurate composition control at the freezing interface. In particular, at the start of freezing, the liquidus composition ( $C_L$ ) in equilibrium with the solid interface must be greater than extremum of the  $(L+\beta)$  phase field, and for stable growth of  $\alpha$ , the average solid composition ( $C_0$ ) must lie such that  $C_L$  does not extend into the  $L+\gamma$  phase field at any time.

## § 2.3 Experimental Procedure

Ingot rods (12.5mm dia x 150mm) of Ti-49at%Al composition were used throughout this study for steady state material feed into the molten zone. The starter rod composition, at the initial freezing interface, was either i) the same as the feed rod, or ii) Al-rich to locally alter liquidus composition  $C_L$ . A single crystal starter seed (of Ti-49at%Al average composition) was also used to promote epitaxial growth. The nominal growth sequence was carried out in discrete steps beginning with 1mm/hr, and ramped to growth rates of up to 8mm/hr. The freezing interface was rotated at 8 rpm, in all the runs, while the melting interface was occasionally counter-rotated at 2-4 rpm to level the thermal profile. This is characteristically different from the prior melting practices employed on this setup.

## § 3 Experimental Results

### § 3.1 similar composition rods

Figure 3 shows generic macrographs of the crystals grown in the current study, where the lamellar plates are always inclined ( $\approx 30-50^\circ$ ) to the crystal growth direction for all runs. The experimental variable is the molten zone hold time prior to solidification traverse, and is 1/4-1/2 hr

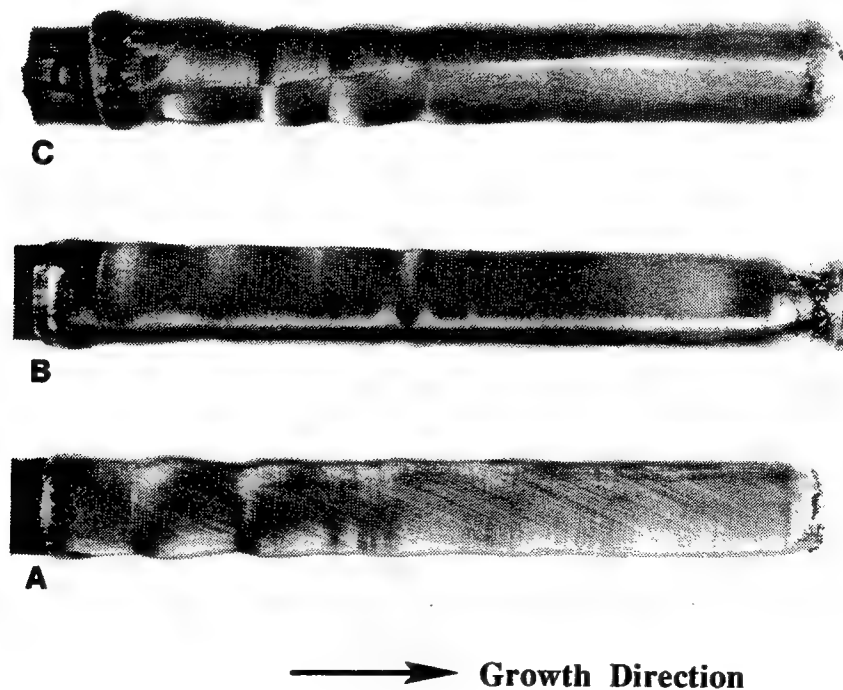


Figure 3. Micrographs of crystals grown using the modified procedure. The initial transients are significantly reduced.

for crystals A, B and about 2 hours for crystal C. Crystals A and B yielded a single crystal cross-section, after a brief initial transient, which is significantly smaller than all previous runs made on this experimental setup [4]. Crystal C (which is typical of a number of runs), grew with a small number of stable orientations in the cross-section. While some competitive grain growth was observed through the length of the processed ingot, it yielded a bi-crystal, with random high angle misorientations across the boundary (arrow in figure 2(c)).

### § 3.2 starter single crystal rods

Single crystal seeds (grown previously Ti-49at%Al rods) were employed as starter rods to promote epitaxial growth of the seed orientation. Results show that in all cases the initially solidified layer produces a myriad of orientations other than the seed orientation, figure 4(a). Thus the seed crystal is, on all occasions, separated from the molten zone and does not propagate into the feed rod. Attempts were made to alleviate this situation by traversing the molten zone into the seed crystal, with no distinct change in results, figure 4(b). This inability to propagate the seed crystal is discussed, with reference to the phase diagram, in the next section.

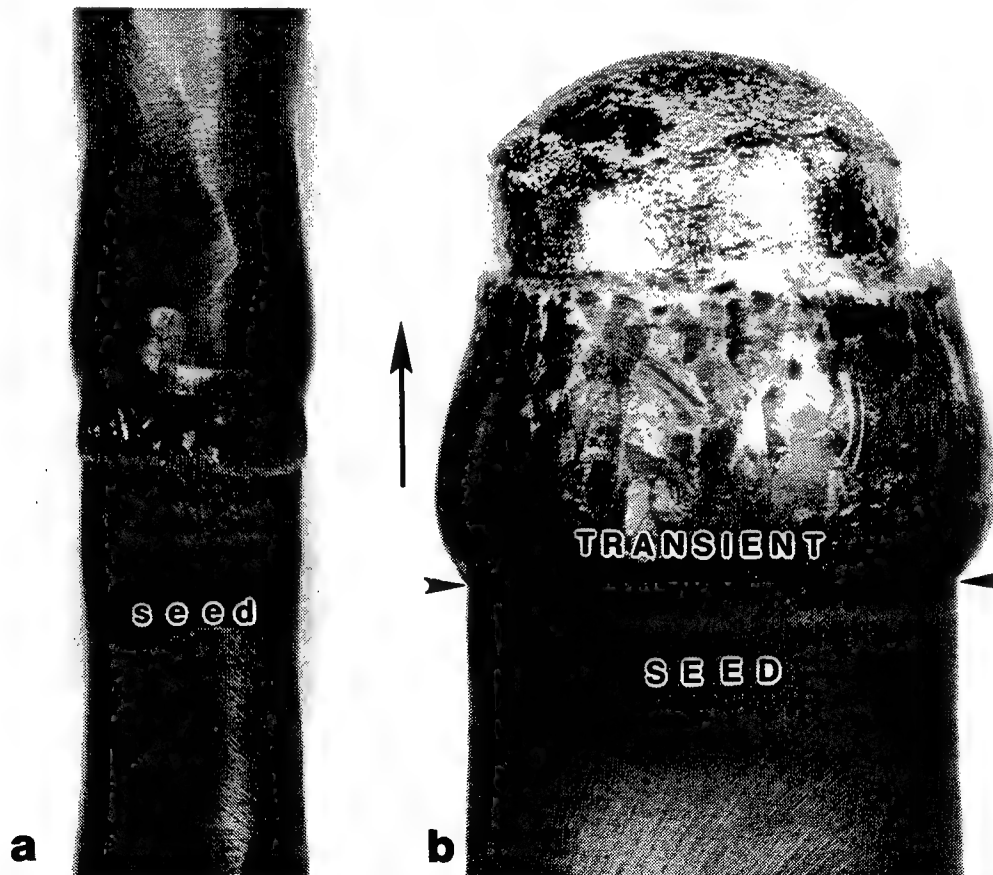


Figure 4. a) Crystal growth (↑) using seed crystals, indicating the inability to epitaxially propagate the seed orientation into the feed rod, b) close-up of initial transient.

### § 3.3 dissimilar composition rods

Starter rods of Ti-52at%Al composition were used in conjunction with a Ti-49at%Al feed rod. The generic nature of the crystals grown is shown in figure 5. Following a brief initial transient, oriented single crystal cross-sections are readily obtained, such that the lamellar plates are aligned predominantly parallel to the crystal growth direction. However, crystal growth was found to be more susceptible to mass transport and control fluctuations. Looking back to the section §1.1, we note that this is the desired orientation for alloy development.

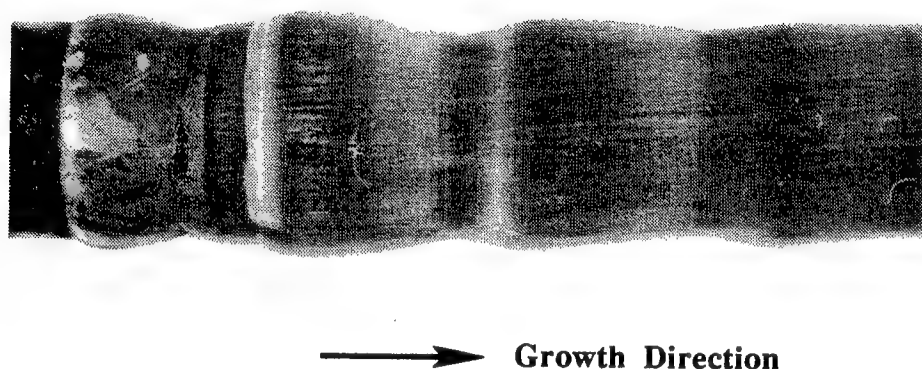


Figure 5. Zone leveled lamellar growth parallel to the crystal growth direction.

## § 4. Discussion

We begin with a description of the solidification phenomenon occurring during zone melting. The experimental results are then discussed in comparison to prior processing practice, and within the framework of i) phase-diagram, and ii) solidification criterion. Changes in processing technique and remedial measures instituted during the course of this study are presented with reference to appropriate experimental results.

### § 4.2 Metallurgical considerations

For the schematic phase diagram for alloy composition  $C_0$ , figure 6(a), zone melting produces a composition transient from  $C_s = kC_0$  through  $C_0$  in the processed bar, where  $k = C_s/C_l$  is the distribution coefficient. As the zone advances, at constant volume, material of concentration,  $C_0$ , enters the zone at melting interface, and solid ( $C_s$ ) of concentration  $kC_l$  leaves the zone at the freezing interface. For  $k < 1$ , the zone accumulates solute as it travels (with a related increase in  $C_s$ ) until the liquid reaches the composition  $C_0/k$ . This represents steady state, as solute mass balance is achieved at the melting and freezing interfaces, yielding a processed bar of uniform  $C_0$  composition, figure 6(c). The solute profile ahead of the freezing interface is given by the equation:

$$C_1/C_0 = 1 + (1/k - 1) \exp [-(R/D)X] \quad (1)$$

where  $C_1$  = solute concentration in the liquid;  $C_0$  = starting solute concentration in the melt;  $k$  = distribution coefficient;  $R$  = growth rate;  $D$  = diffusion coefficient of the solute in the liquid; and  $X$  = distance measured from the interface into the liquid. Corresponding to this variation of composition in the liquid, the liquidus temperature, see figure 6(b), varies as [3]:

$$T_1 = T_0 - mC_0 \{ 1 + (1/k - 1) \exp [-(R/D)X] \} \quad (2)$$

where  $T_0$  = melting point of pure solvent metal;  $T_1$  = liquidus temperature;  $m$  = liquidus slope. This liquidus temperature, if higher than the actual temperature in the liquid, as imposed by the freezing conditions, gives rise to constitutional supercooling in a limited region, see figure 6(b). This constitutes an instability in the planar solidification front, essential for single crystal growth. Single crystal growth, is effected by maintaining a sufficiently high  $G/R$ , where  $G$  is given by equation 2.

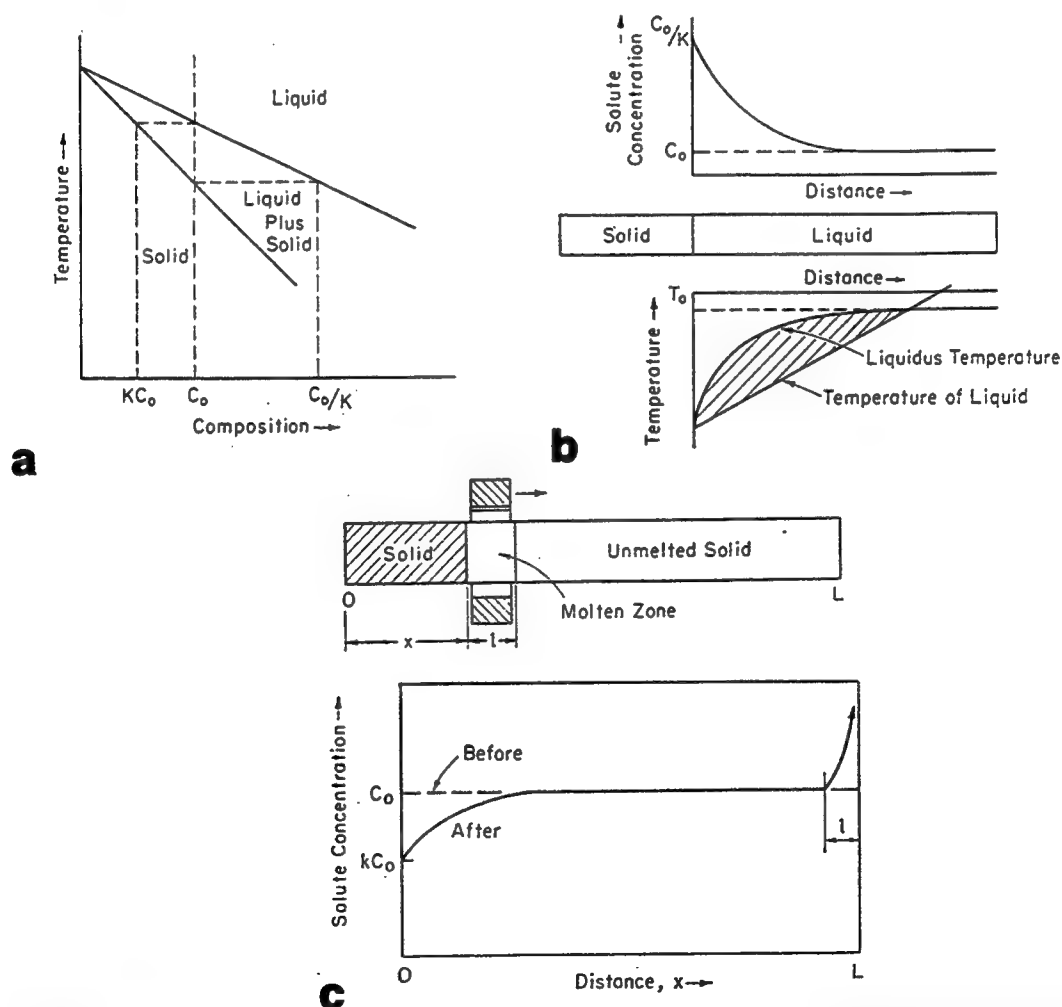


Figure 6. a) Schematic binary phase diagram illustrating, b) solute profiles at the freezing interface, and constitutional supercooling that destabilize stable crystal growth, and c) resulting compositional profile along the length of the processed bar

## § 4.2 Prior processing practices

A brief synopsis of this prior melting practice is as follows; the melts were made with synchronous rotations of the starter and feed rods through the molten zone. Since the field coil does not provide liquid zone levitation based confinement, it is primarily confined by surface tension effects. This requires that the molten zone be maintained at low superheats (i.e., with very little turbulence and mixing, to avoid spilling) and consequently there is a lower temperature gradient in the liquid. The lack of turbulent mixing, and a large boundary layer adjacent to the freezing interface, ensures that the composition  $C_1$  approaches steady state very gradually, see figure 6(c). This has the net result of creating large initial transient regions, often half the length of the processed bar [4], with significant compositional variations along the length of the crystal. Additionally, a low  $G$  necessitates that lower overall growth rates ( $R$ ) be employed (i.e., with reduced throughput) to maintain planar front solidification, necessary for single crystal growth. Changes in processing variables and remedial measures employed during the course of this study are presented in the following paragraphs, with reference to their appropriate experimental sections.

## § 4.4 similar composition rods

The temperature gradient in the liquid ( $G$ ), adjacent to the freezing interface, exists over a boundary layer thickness ( $\delta$ ) extending from the interface (where there is no mixing whatsoever) into the liquid, figure 6(b). We note while  $G$  is inversely proportional to  $\delta$ , the steep liquidus temperature slope at the freezing interface, remains virtually unaffected. Thus,  $\delta$  can be reduced (to increase  $G$ ) by i) increasing fluidity by superheating, ii) mixing in the liquid, and iii) by rotating the freezing interface. Both i) and ii) can be used only minimally as there are limits, as imposed by current field coil design, to the level of turbulence sustained by the liquid zone. Thus, only rotations of 8rpm were employed, which significantly reduce the initial transient region to  $< 15\text{mm}$  consistently. Reducing  $\delta$  also allows the liquidus composition at the interface, to saturate to the steady state composition  $C_0/k$  earlier, thereby simultaneously reducing the compositional transient, see figure 6(c), along the length of the processes bar. Note that rotations employed in this study are quite conservative, primarily selected to avoid spilling the surface tension confined liquid zone. Rotations of up to 100rpm have been employed elsewhere [5] to sustain planar front growth, at significantly higher ( $\approx 12\text{mm/hr}$ ) growth rates, for 25mm diameter bars.

A single crystal cross-section was more likely to be obtained when initiating growth soon after stabilizing the molten zone. Long hold times, generally produced a large number of orientations, all equally suited for growth. This invariably produced bi-crystals, see figure 3(c) of with a variety of

misorientations, and this condition must be avoided. Incidentally, these observations suggest that single crystal growth may be promoted by rapid growth in the initial transient, such that a favorable orientation may grow large at the expense of others.

#### § 4.5 single crystal seed

Efforts to initiate epitaxial growth of the seed orientation were not successful, in the present case. Any successful seeding operation, requires that the liquid composition enter the  $L+\alpha$  phase field. Because of the uncertainties with the high temperature regime of the phase diagram, particularly when dealing with a molten zone of unknown interstitial impurity, it is unclear whether Ti-49at%Al liquidus intersects the  $L+\alpha$  phase field, and traversing through the  $L+\beta$  phase field nullifies the seed crystal orientation. However, the more severe shortcoming is that the initial seed crystal underwent a long initial transient, and consequently the exact composition at the seed interface was unknown, but is most likely to be less than Ti-49at%Al, see figure 6(c). Since the initial liquidus composition is determined by the seed composition, this increases the possibility of  $\beta$  nucleation thereby negating the seed orientation.

#### § 4.6 dissimilar composition rods

The initial transients invariably introduce compositional uncertainties along the processed ingot length, see figure 6. Processing schemes such as i) increasing superheat, ii) mixing in the liquid, and iii) by rotating the freezing interface, can be employed to reduce the length over which these compositional transients occur, as demonstrated in figure 3, but they do not affect the magnitude of the concentration variation ( $\Delta C$ ). An Al-rich starter rod was used for zone leveling, and to promote  $L \rightarrow \alpha$  solidification. This zone leveling procedure increases the local liquidus composition such that the resulting compositional transient ( $\Delta C$ )  $<$  ( $C_s - C_o$ ). An ideal situation would be start with a liquidus composition ( $C_L$ ) =  $C_o/k$  which results in no transient, as shown in figure 7.

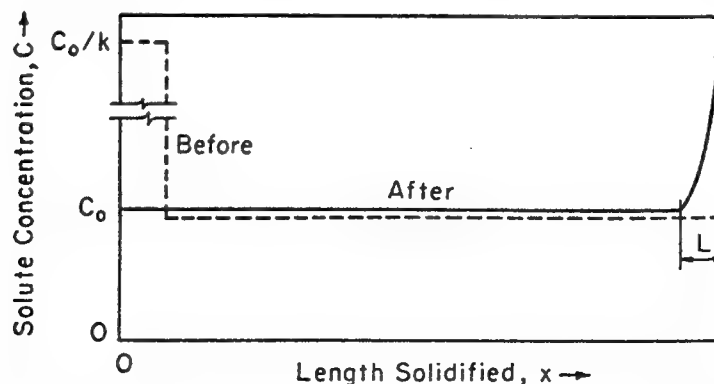


Figure 7. Schematic of the zone leveling process designed to eliminate compositional transients along the length of the processed bar.

However, the critical reason for employing Al-rich starter bar is to ensure that solidification follows the  $L \rightarrow \alpha$  and not  $L \rightarrow \beta$  path and promote preferred growth directions (i.e., those contained in the basal plane) in the  $\alpha$ -hcp phase. Recall that the driving objective here is to obtain oriented  $\alpha$ , the precursor phase to oriented lamellar material. In the results presented here, possible  $\langle 10\bar{1}0 \rangle_\alpha$  growth along the length of the bar produces the desired lamellar orientation.

## Summary and Conclusions

Critical parameters for growth of oriented single crystals of lamellar TiAl were evaluated. Desired oriented growth was obtained by ensuring primary  $L \rightarrow \alpha$  solidification, and avoiding the additional  $L \rightarrow \beta$  step in the solidification path. Direct 'seeding' options must be reevaluated, within the context of ensuring that the seed composition lies within the  $L + \alpha$  phase field.

## Acknowledgments

The author thanks Mike Scott for his help and unyielding patience with the crystal growth apparatus, and thanks Dr. Dennis Dimiduk and Dr. Madan Mendiratta for their valuable suggestions and comments during the course of this study. Financial support from AFOSR contract # F49620-93-C-0063 is acknowledged

## References

1. R.D. Reviere, B.F. Oliver and D.D. Bruns, 1989, Mat. Manufacturing Proc., **4**, 103.
2. B.F. Oliver and B.K. Kad, 1991, J. Less Common Metals, **168**, 81.
3. W. G. Pfann, 1958 "Liquid Metals and Solidification" ASM Conf. Proc., p.283.
4. D. M. Dimiduk and M.J. Scott, *private communication*
5. B.F. Oliver, B.Y. Huang and W.C. Oliver, 1988, Scripta Met., **22**, 1405.



# **A Study of Super Capacitor Applications**

**Marian K. Kazimierczuk  
Professor  
Electrical Engineering Department**

**Wright State University  
Dayton, OH**

**Final Report for:  
Summer Faculty Research Program  
Wright Laboratory**

**Sponsored by:  
Air Force Office of Scientific Research  
Bolling AFB, Washington, DC**

**and**

**Wright Laboratory**

**August 1995**

# A STUDY OF SUPER CAPACITOR APPLICATIONS

Marian K. Kazimierczuk  
Professor  
Department of Electrical Engineering  
Wright State University  
Dayton, OH 45435

*Abstract* — The feasibility of the use of a super capacitor and DC/DC converter to improve the regulation of the bus voltage of an aircraft was studied. A boost DC/DC converter was used to transfer the energy from the capacitor to the dc bus voltage level. A buck DC/DC converter was used to transfer energy from the dc bus to the capacitor. Experimental results indicate that a super capacitor and DC/DC converter can be used to improve the regulation of the bus voltage of distributed power systems.

# A STUDY OF SUPER CAPACITOR APPLICATIONS

Marian K. Kazimierczuk  
Professor  
Department of Electrical Engineering  
Wright State University  
Dayton, OH 45435

## I. INTRODUCTION

Recently, capacitors with large capacitances have been developed [1]–[4]. These capacitors are called *super capacitors*. The range of the capacitances is on the order of 1 to 500 F. However, The voltage ratings of the super capacitors is on the order of only 2.5 to 15 V.

The principal operation of super capacitors was first investigated by Helmholtz [5] in 1879. One electrode of a super capacitor is made of carbon and the other is made of a liquid electrolyte. When a positive voltage is applied to the carbon layer and the negative voltage applied to the liquid electrolyte, a thin dielectric layer is established. The plate area is extremely large, on the order of 1000 m<sup>2</sup>/g, because of the porous surface of the carbon. The thickness of the dielectric layer is extremely small – on the order of 1 nm. As a result, a high volumetric efficiency is obtained. Capacitance densities on the order of 30 F/g of carbon and 1 F/cm<sup>3</sup> are achievable.

This high capacitance permits the storage of large amounts of energy which leads to a large number of new applications. A power electronics circuit can be used to condition the output voltage and current of a super capacitor. The combination of super capacitor and power electronic circuit can be called a *super battery*.

An electric aircraft contains a distributed power system. Its nominal dc line voltage is 270 V. This voltage can vary from 250 to 280 V for steady state. When actuators are activated, a larger

load is added to the system. This causes an increased current to flow through the distributed line resistance and inductance thereby reducing the voltage seen at the load. Super capacitors can be used to improve the voltage regulation at the point of load in the aircraft distributed power systems. The objective of this research was to investigate the feasibility of improving the voltage regulation of a distributed power system.

## **II. LIMITATIONS OF BATTERIES**

1. Environmental hazard
2. Safety problems
3. Maintenance cost
4. Slow charging
5. Limited number of charge cycles and life
6. Memory problems in some batteries such as NiCd
7. Complicated charging circuits
8. Need for continuous replacement

## **III. FEATURES OF SUPER CAPACITORS**

1. Very high energy density (20 times that of conventional capacitors)
2. Ultra long life. They can be fully charged and discharged more than 100,000 times with an expected operating lifetime longer than 25 years. Unlike batteries, super capacitors have no parasitic chemical reactions.
3. Some super capacitors are nonpolar.
4. Some super capacitors do not require current limiting resistors or over-voltage protection.
5. Maintenance free
6. Safer than batteries – super capacitors will not explode if short circuited.
7. Volume energy density 9 Wh/L
8. Weight energy density 4 Wh/kg
9. Volume power density 900 Wh/L
10. Weight power density 400 Wh/Kg

## IV. APPLICATIONS OF CAPACITORS

1. Memory backup power supplies for computers, timers, and other electronic equipment such as security systems and programmable controllers
2. Power supplies for smoke detectors
3. Power supplies for emergency lights
4. Starters, ignitors, and actuators such as ignition systems for automobiles
5. Emergency power sources for aircraft
6. Voltage regulation in distributed systems such as aircraft power systems under variable load conditions
7. Source of large instantaneous power levels such as for aircraft actuators
8. Voltage regulation of systems with switched loads such as phase radars
9. Active power filters

## V. EXPERIMENTAL RESULTS OF IMPROVED VOLTAGE REGULATION USING SUPER CAPACITORS

### A. Switched Load

A circuit that simulates the variations of voltage in a switched load is needed. The voltage changes normally in a continuous manner. The worst case occurs when the change takes place instantaneously. Fig. 1(a) shows a switched load circuit. It consists of a regulated dc voltage  $V_I$ , load resistors  $R$ , and a power MOSFET. When the MOSFET is off, the output voltage is  $V_I/2$ . On the other hand, when the switch is on, the output voltage is  $V_I/3$ . If the MOSFET is switched periodically at 100 Hz, the load voltage  $V_O$  will jump from  $V_I/3$  to  $V_I/2$  and vice versa.

### B. Test Circuit for Line Voltage Regulation Improvement

Fig. 1(b) shows a test circuit for line voltage regulation improvement. It consists of a super capacitor  $C$ , a step-up DC/DC converter, a switched load, and a dc bus voltage  $V$ . When the MOSFET is turned on, the effective load resistance is a parallel combination of  $R_2$  and  $R_3$ . This causes the output voltage  $V_O$  to decrease. The DC/DC converter transfers the energy from the super capacitor to the switched load during this time interval. As a result, the load receives more

current and the voltage is increased. Therefore, the load regulation of the bus voltage is improved. On the other hand, when the load resistance is too light, the bus voltage increases and the energy is transferred from the load to the super capacitor.

In the case of unidirectional step-up converter, the lower voltage level is increased close to the upper level. The super capacitor can improve voltage regulation for a limited length of time. When the super capacitor is discharged, its voltage decreases, and the voltage transfer function of the converter increases to its maximum value. At this time, the converter stops regulating the bus voltage.

Fig. 2 shows a bidirectional DC/DC converter, super capacitor  $C$ , and a load resistance  $R_L$ . During the discharging mode of the super capacitor, the circuit acts like a boost converter and transfers the energy from the super capacitor to the bus. Conversely, during the charging mode, the circuit acts like a buck converter and transfers the energy from the bus to the super capacitor.

### C. Test Results

The super capacitor was built using four Panasonic 3.3 F/5.5 V super capacitors. Two pairs of capacitors were connected in series which in turn were connected in parallel to achieve a 3.3 F/11 V super capacitor bank.

A step-up boost circuit was built using an International Rectifier IRF530 power MOSFET, a Motorola MUR860 ultrafast recovery diode, an inductor  $L = 183.8 \mu H$ , and a filter capacitor  $C = 47 \mu F$ . The output voltage  $V_L$  was 20 V. The switching frequency of the converter was 100 kHz.

The switched load was built using  $R_1 = R_2 = R_3 = 50 \Omega$  and an IRF530 power MOSFET. The switching frequency of the switched load was 100 Hz.

Fig. 3 shows the waveforms of the line voltage in the circuit of Fig. 2(b) with the disturbance of the line voltage caused by a switched load and improvement due to the super capacitor and boost DC/DC converter at time  $t = 0, 10, 25, 35, \text{ and } 40$  s. It can be seen that the peak-to-peak bus voltage is very low at  $t = 0$ . As  $t$  increases, the regulation improvement deteriorates. At  $t = 40$  s, the converter quit regulating the bus voltage and there was no longer any improvement in the line voltage.

Fig. 4 shows the peak-to-peak ripple of the line voltage in the super capacitor discharge mode with a switched load. Fig. 4(a) shows the output ripple voltage  $V_{opp}$  versus time  $t$  and Fig. 4(b) shows the output ripple voltage  $V_{opp}$  versus super capacitor voltage  $V_C$ .

#### *D. Fixed Load*

Fig. 5(a) shows the output voltage  $V_O$  of the super capacitor versus time  $t$  with a fixed load resistance  $R_L = 100 \Omega$ . Fig. 5(b) shows the output voltage  $V_O$  versus super capacitor voltage  $V_C$  with a fixed load resistance  $R_L = 100 \Omega$ .

### VI. SUPER CAPACITOR CHARACTERISTICS

Fig. 6(a) shows the measured plots of capacitance  $C_S$  and equivalent series resistance  $R_S$  versus frequency  $f$  for Panasonic 3.3 F/5.5 V super capacitor. The capacitance decreases with frequency substantially as does the equivalent series resistance. The self resonant frequency occurs at approximately 290 kHz. The magnitude  $|Z|$  and phase  $\phi$  of the super capacitor versus frequency are shown in Fig. 6(b). The equipment used did not all the measurement of the capacitor below 100 Hz.

### VIII. CONCLUSIONS

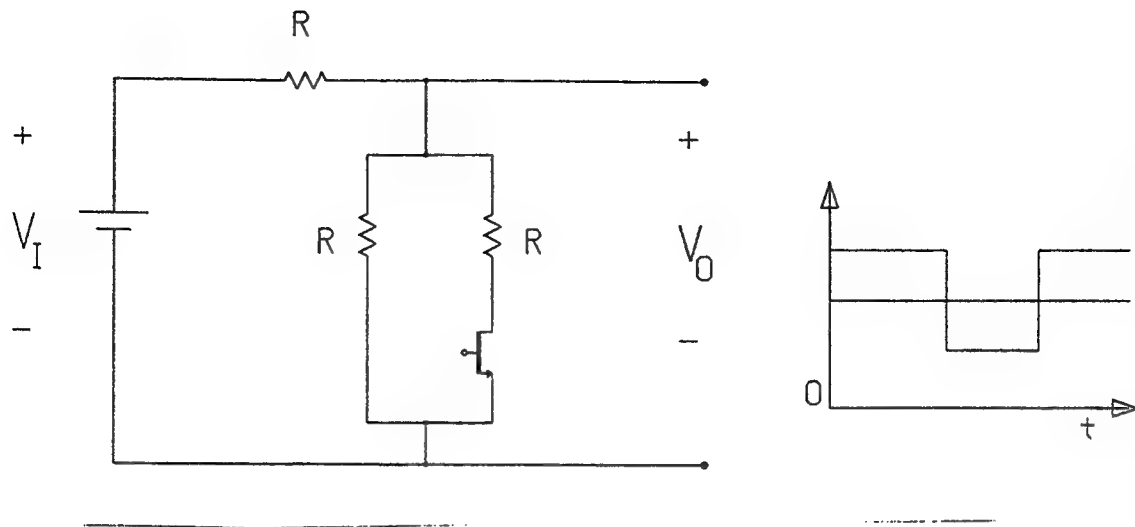
Experimental evidence has shown that super capacitors can be used to improve the line voltage regulation. In the demonstrated experiment, the line voltage regulation was improved for approximately 35 s. Super capacitors can be also used to provide emergency power at fix loads.

### References

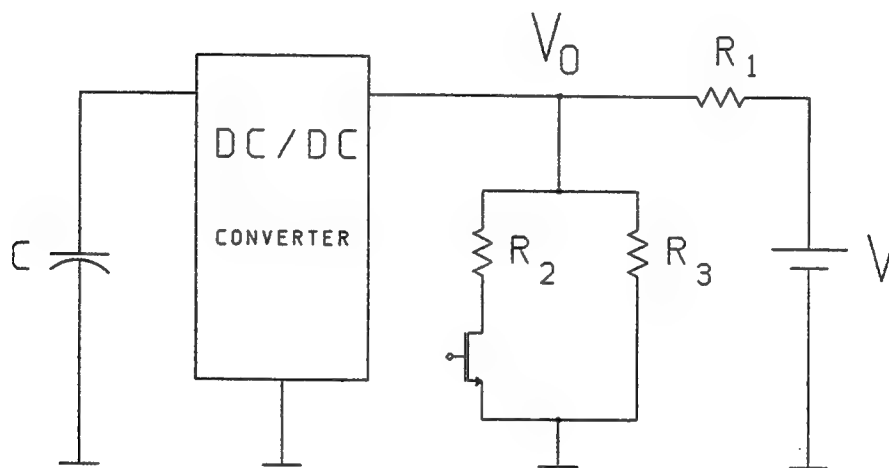
- [1] The Evans Capattery Data Sheets.
- [2] The Evans Capattery, The Next Generation in Double-Layer Capacitors, Technical Papers 1989-1994.
- [3] Cesiwid Maxcap Double Layer Capacitors, Product Information and Application Data.
- [4] The Fourth International Seminar on Double Layer Capacitors and Similar Energy Storage Devices, Boca Raton, FL, December 12-14, 1994.

[5] H. L. von Helmholtz, "Wied. Ann.," 7, 33, 1879.





(a)



(b)

Figure 1: (a) Circuit diagram of switched load. (b) Line regulation improvement test circuit consisting of super capacitor  $C$ , DC/DC converter, switched load, and bus voltage  $V$ .

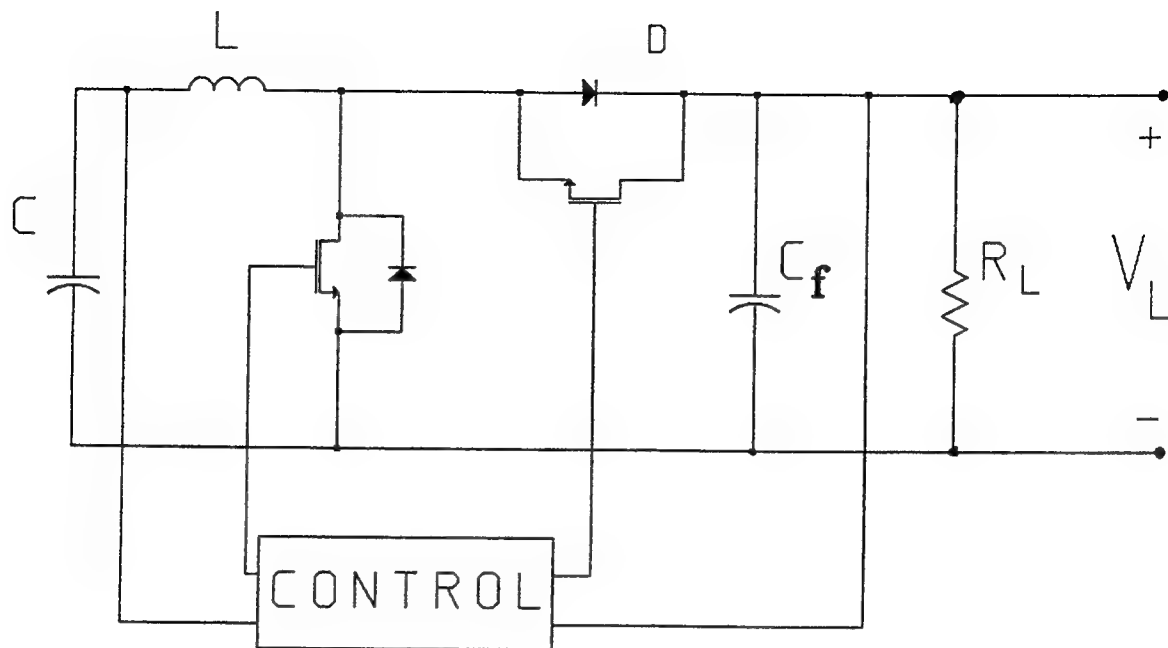


Figure 2: Circuit diagram of bidirectional DC/DC converter, super capacitor  $C$ , and load  $R_L$ .

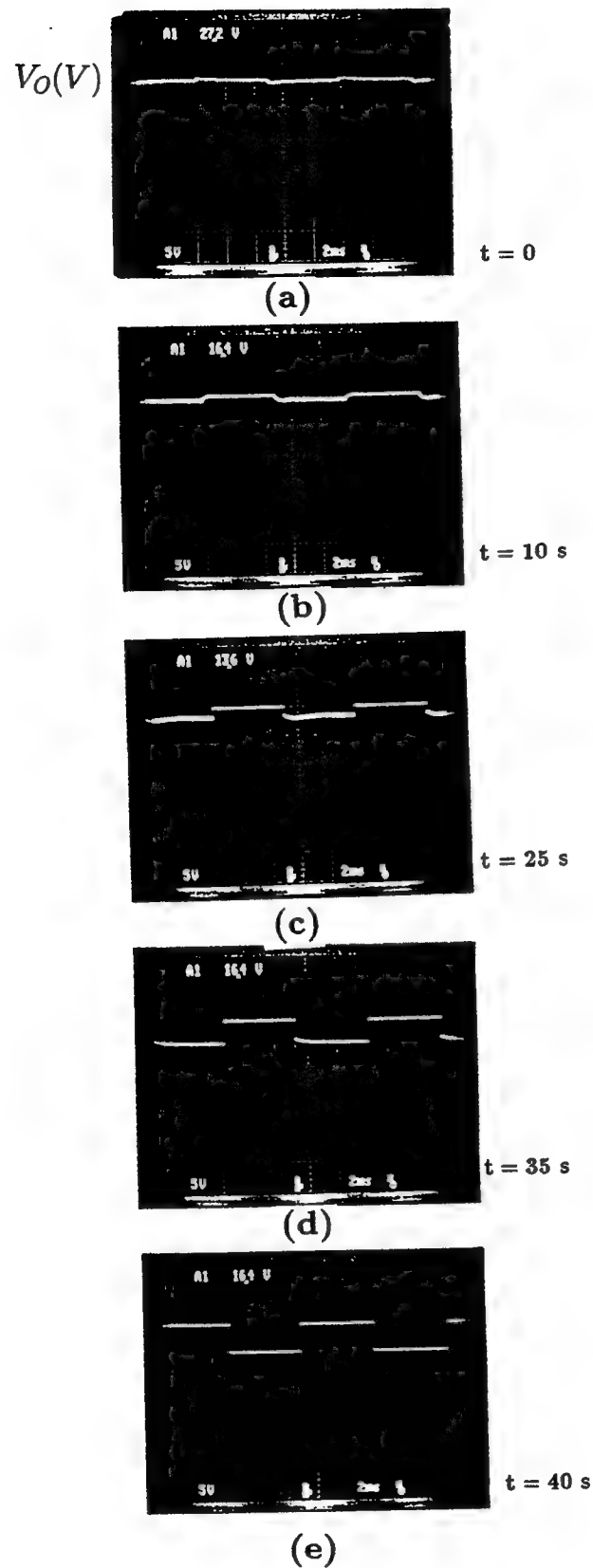
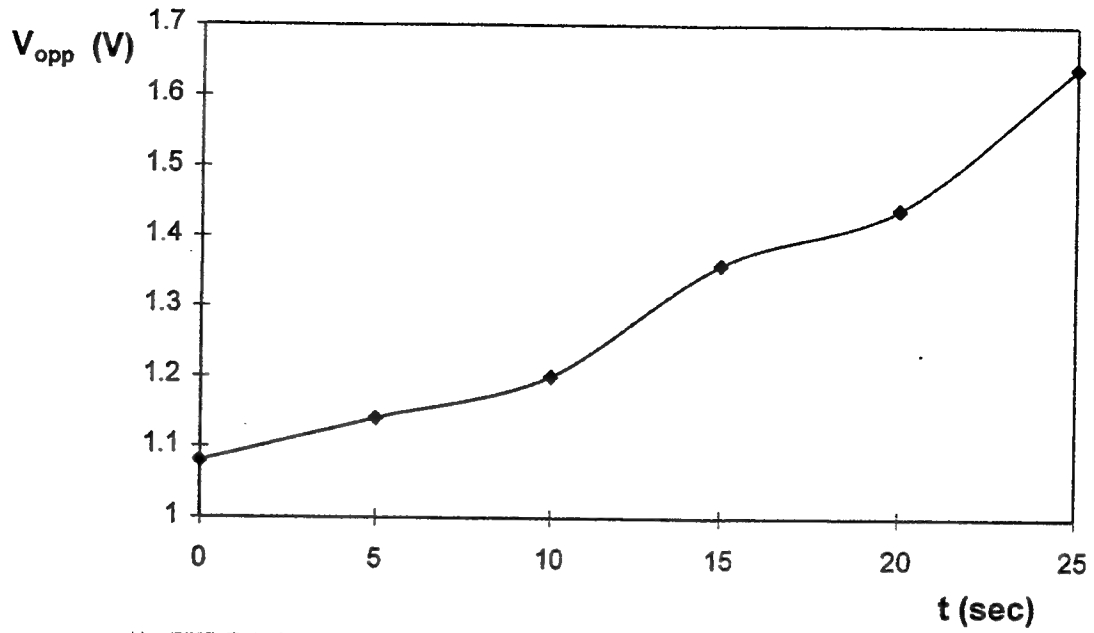
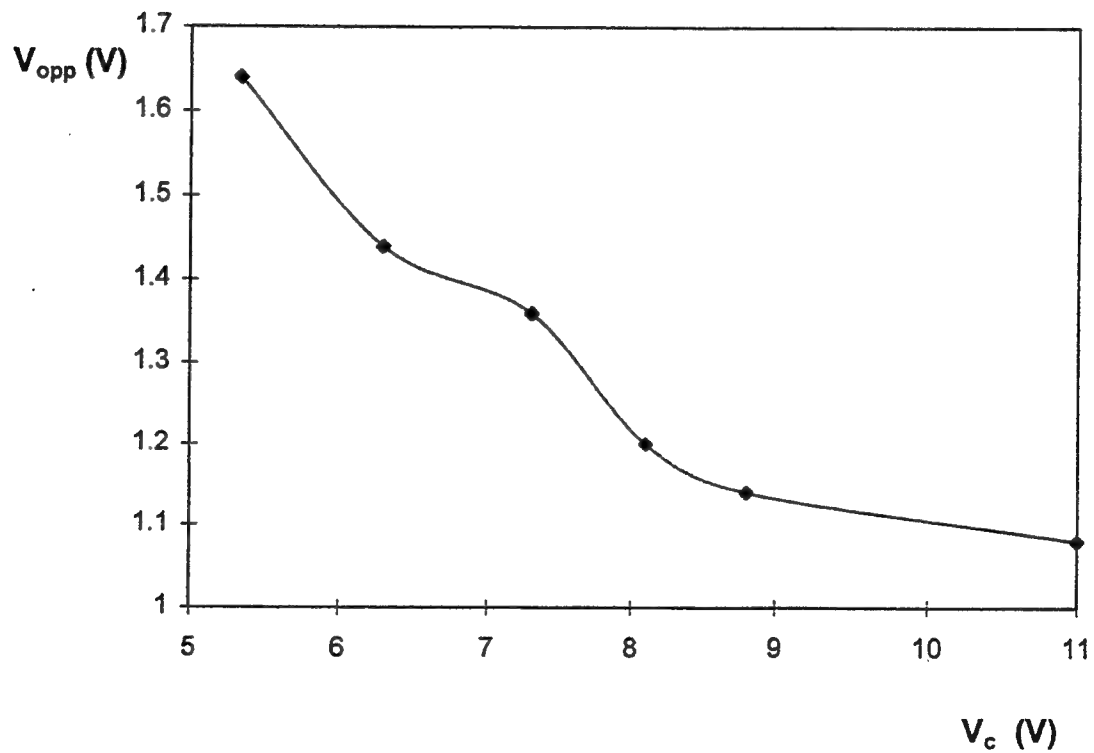


Figure 3: Waveforms of the line voltage in the circuit of Fig. 2(b) with the disturbance of the line caused by a switched load and improvement due to the super capacitor and boost DC/DC converter. (a) At  $t = 0$ . (b) At  $t = 10$  s. (c) At  $t = 25$  s. (d) At  $t = 35$  s. (e) At  $t = 40$  s. Horizontal: 2 ms/div. Vertical: 5 V/div.

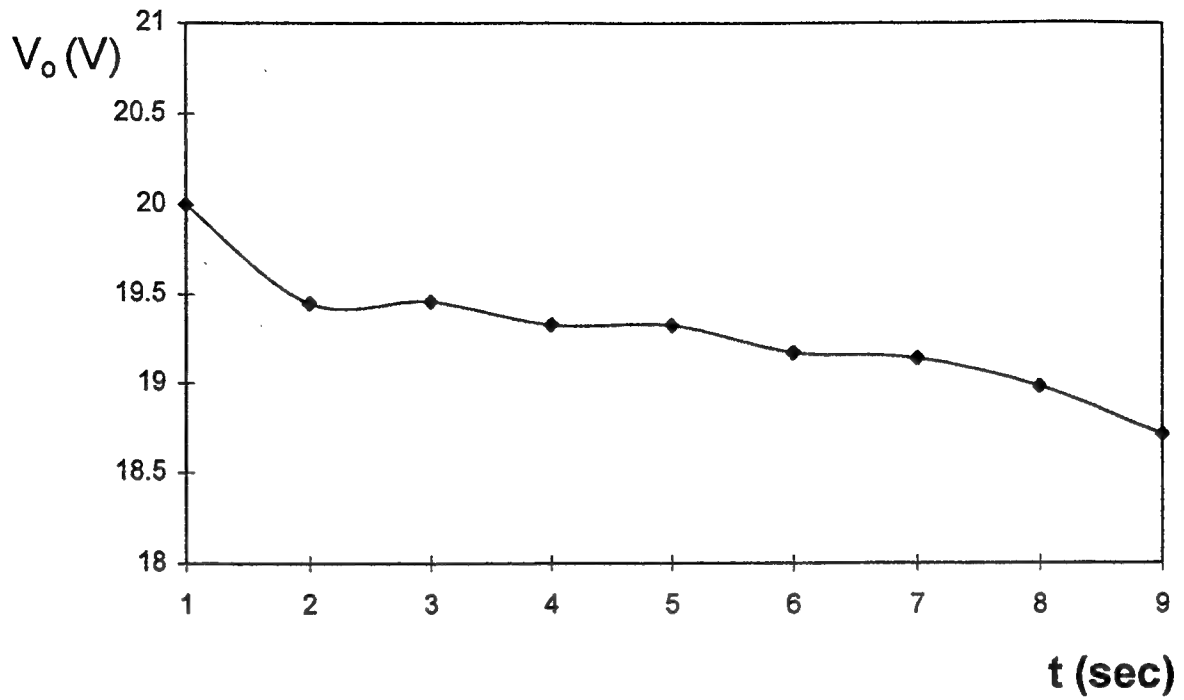


(a)

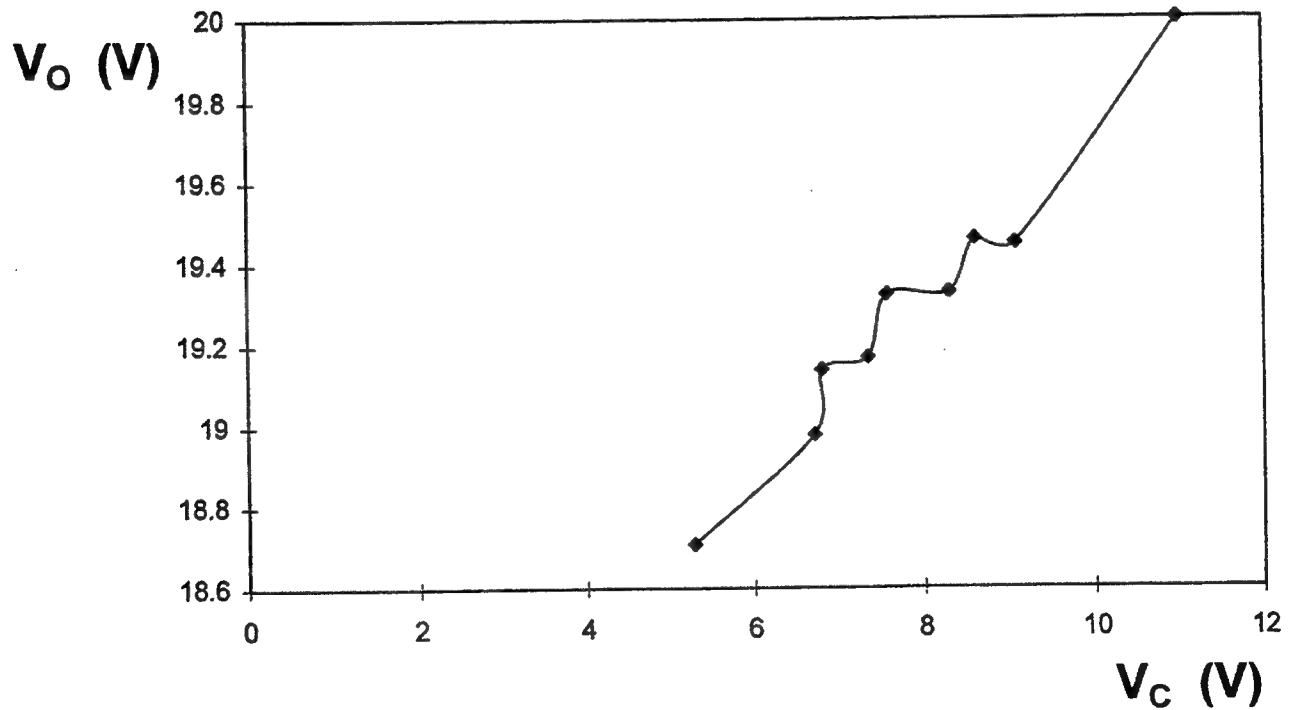


(b)

Figure 4: Peak-to-peak ripple of the line voltage in the super capacitor discharge mode with switched load. (a) Output ripple voltage  $V_{opp}$  versus time  $t$ . (b) Output ripple voltage  $V_{opp}$  versus super capacitor voltage  $V_c$ .

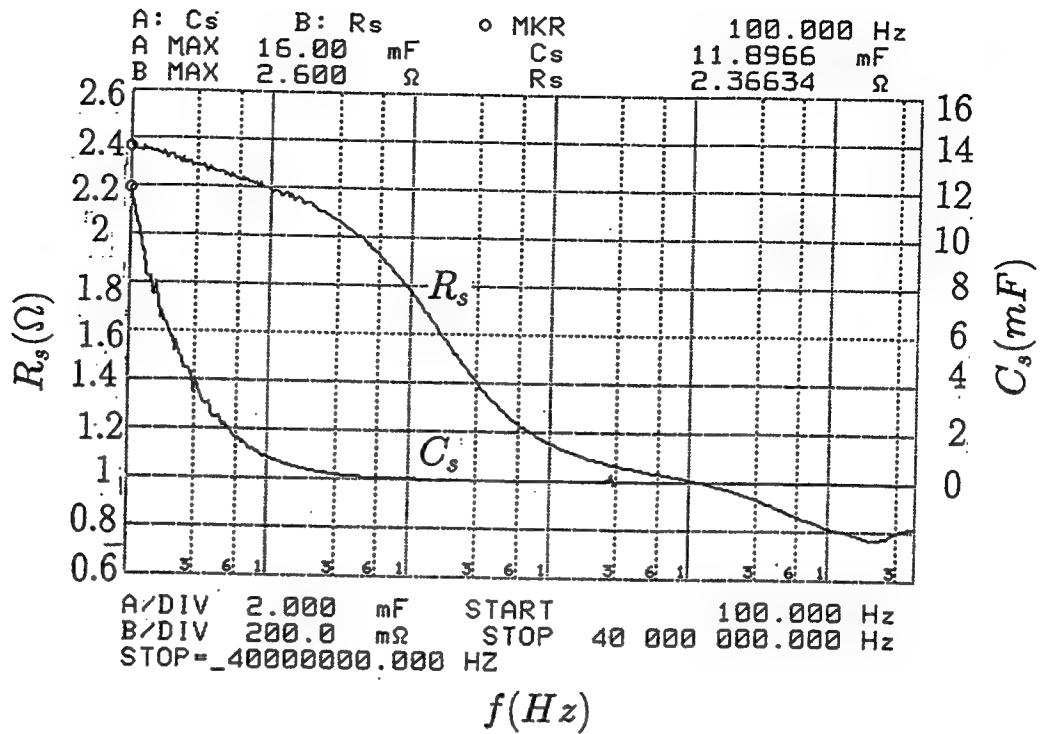


(a)

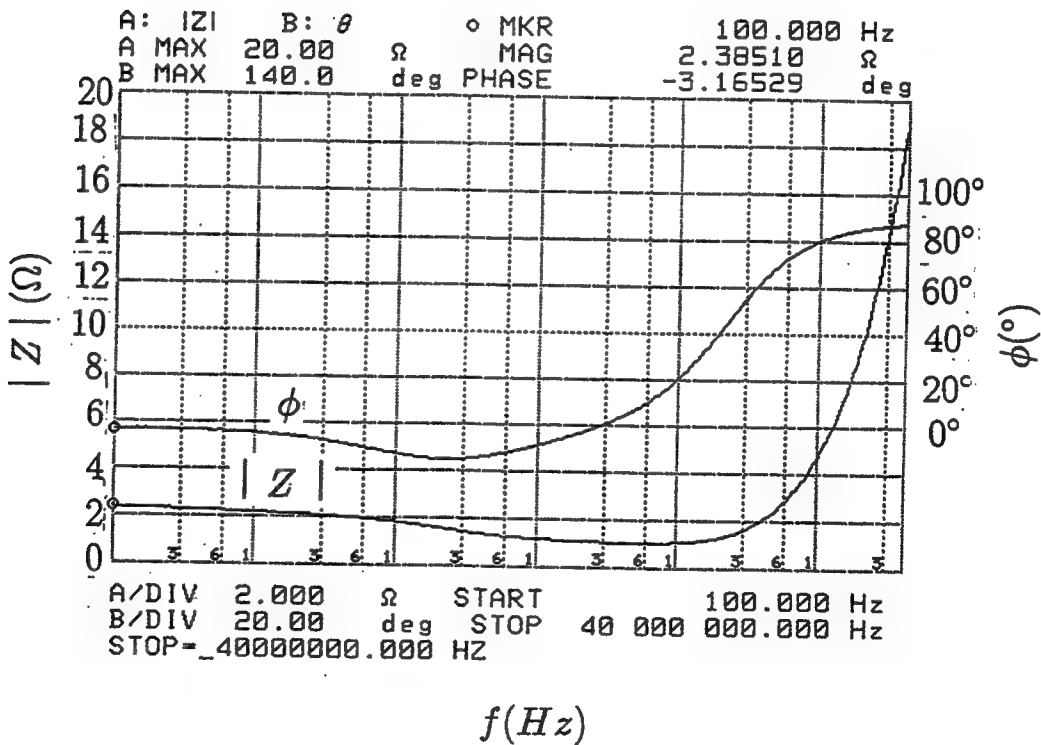


(b)

Figure 5: Output voltage in the super capacitor discharge mode with fixed load. (a) Output voltage  $V_O$  versus time  $t$ . (b) Output voltage  $V_O$  versus super capacitor voltage  $V_C$ .



(a)



(b)

Figure 6: Super capacitor characteristics. (a) Capacitance  $C_s$  and equivalent series resistance  $R_s$  versus frequency  $f$ . (b) Magnitude  $|Z|$  and phase  $\phi$  of the super capacitor impedance versus frequency  $f$ .

**PART I: EXCAVATOR-MANIPULATOR SYSTEM FOR NEUTRALIZING UNEXPLODED  
ORDNANCE**

**PART II: ADAPTIVE SELF-TUNING CONTROL OF EXCAVATORS**

**A. J. Koivo  
Professor  
School of Electrical and Computer Engineering**

**Purdue University  
West Lafayette, IN**

**Final Report for:  
Summer Faculty Research Program  
Wright Laboratory**

**Sponsored by:  
Air Force Office of Scientific Research  
Bolling AFB, Washington, DC**

**and**

**Wright Laboratory**

**August 1995**

## **PART I: EXCAVATOR-MANIPULATOR SYSTEM FOR NEUTRALIZING UNEXPLODED ORDNANCE**

## **PART II: ADAPTIVE SELF-TUNING CONTROL OF EXCAVATORS**

**A. J. Koivo**  
Professor  
School of Electrical and Computer Engineering  
Purdue University

### **Abstract**

**Part I:** Certain areas for example on beaches can contain unexploded ordnance which can be very dangerous and life-threatening. Therefore the neutralization of such ordnance is essential to make the land useful. After exposing unexploded ordnance, a human usually attaches a rocket-wrench fixture to the fuze of the detonating mechanism. Cartridges attached to this fixture can then be fired to remove the fuze from the ordnance. Since it is dangerous and potentially life-threatening to work in the close neighborhood of unexploded ordnance, a robotic system is designed here to perform the foregoing task of placing the specific fixture around the ordnance.

The mechanical system consists of two robotic manipulators. They can be mounted on the same platform which is attached to an excavator as an end-effector. The excavator will transfer the robotic system close to the ordnance. Then, the motion of the dual-arm system can be guided by a teleoperator so that it will transfer the rocket-wrench to and around the fuze, where the jaws of the wrench assembly will be tightened for grasping the fuze. To control the dual-arm manipulator, an architecture for the teleoperated system is presented.

**Part II:** To automate the digging operations of an excavator, an adaptive feedback controller is designed here so that the bucket pose tracks a specified trajectory. The controller is determined by minimizing the expected value of the squared tracking errors and the consumed energy subject to a difference equation constraint. The latter represents the input-output relations of the excavator dynamics. In this time-series difference equation, the estimated parameter values are used. The resulting controller is in a feedback form, and the gains depend on the last (best) parameter estimates. Thus, when a new set of parameter estimates are calculated on the basis of the latest measurements, the controller gains are changed accordingly (adaptation). The performance of the designed adaptive self-tuning controller is illustrated by simulations.



# **PART I: EXCAVATOR-MANIPULATOR SYSTEM FOR NEUTRALIZING UNEXPLODED ORDNANCE**

**A. J. Koivo**

## **1. INTRODUCTION**

Unexploded ordnance are known to exist in certain areas which can be quite large. Whether ordnance are partially or completely underground, for example, in sand on a beach, the first task usually after detection is to determine the location, and specifically the pose (the position and orientation) of the ordnance. Systems have been developed, implemented and tested for detecting and locating underground ordnance. After the localization of the unexploded ordnance, the task is to neutralize them. The neutralization is usually accomplished by first exposing the ordnance so that the fuze (or the detonation mechanism) can be seen. Then, the fuze is usually removed with manual help to make the ordnance harmless.

The removal of the fuze is presently performed with the help of a human. However, a human working at the site of the unexploded ordnance is dangerous, and can result in accidents and deaths of humans. Therefore, it would be desirable to have a mechanical system to perform the task of removing the fuze from the ordnance without having a human in the neighborhood of the ordnance at anytime during the operation. This mechanical system could be controlled remotely by a human, or autonomously controlled using appropriate computer control with sensory information.

Recent advances on the control of robotic manipulators and dexterous manipulations do offer a possible solution. The task of removing a fuze mechanism from an exposed unexploded ordnance does require fine motion control and dexterity. The technology to dispose the fuze mechanism by robotic manipulators is available. In the sequel, we will outline a possible approach to design and operate a robotic system that will remove an exposed fuze mechanism from an unexploded ordnance.

## **2. TASK DESCRIPTION**

It is assumed that the fuze for the detonation of the ordnance has been exposed and is visible. Moreover, the fuze mechanism is assumed to be located at the rear or in front of the ordnance. In removing the fuze of the foregoing type, a fixture commonly used is a rocket-wrench assembly. It is proposed here that a robotic system will be used to transfer and attach the rocket-wrench fixture tightly to the fuze mechanism. Then by firing the cartridges on the rocket assembly, the fuze of the ordnance can be removed. Thus, the bomb is made harmless. An excavator can be used to move the robotic system to and from the ordnance.

The rocket-wrench assembly has two bolts on the opposite sides to tighten the jaw assembly around the fuze (wrench-type function). The turning of the bolts (one at a time) tightens

the jaw assembly around the fuze. It can be performed manually, or by a robotic hand rotation, or by an attached servomotor. It is important that the jaw assembly is so tightened that the center of gravity of the rocket-wrench assembly lies on the geometric axis of the fuze so that the rocket-wrench at firing the cartridges will rotate about this geometric axis without damaging the threads of the fuze.

The components of a commonly used rocket-wrench assembly has the following approximate weights: jaw assembly 4 lbs 13.9 oz, rocket assembly 6 lbs 6 oz, two cartridges each 2.30 oz, which contain powder 0.0336 lbs each. The total weight of the rocket-wrench assembly is (about) 11 lbs 9.6 oz, or equivalently 5.260 kg. A dual arm manipulator system is proposed for the task in view of the weight of the total load and the necessary dexterous manipulation [1]. The two aforementioned bolts could be grasped by the plate-fingers of the two robotic manipulators. This grasp arrangements will allow (i) the carrying and the transfer of the rocket-wrench fixture by the robot manipulator system, and (ii) the tightening of the jaw assembly symmetrically.

### 3. OVERALL SYSTEM

The overall system consists of an excavator which can carry a robotic manipulator system to the neighborhood of the ordnance so that the fuze will be in the workspace of the robotic system. If one considers the excavator as a macro-robot and the end-effector, i.e., the terminal arm system as a mini-manipulator system, then the overall system can be called a macro-mini manipulator system. The macro-manipulator to be used can be a standard excavator such as CAT 320L of Caterpillar. A platform on which the manipulator system is mounted will be the end-effector of the excavator, or attached to it. The mechanical attachment mechanism of the platform to the excavator can be similar to the one used for a bucket. The focus of the following discussion will be on the robotic manipulator system.

It will be assumed that the removal of the fuze mechanism from the ordnance will be performed by means of the rocket-wrench which is currently used by humans performing the task in question. After the rocket assembly is attached to the wrench assembly, the approximate weight of the rocket-wrench assembly with two cartridges for rocket action is 5.260 kg. The robotic manipulator system must carry this load to a place where it can grasp the fuze capsule. If the rocket-wrench load is to be carried and manipulated by a single robotic manipulator, this load is heavy and difficult to handle. Moreover, a single robotic manipulator with a load carrying capacity of at least 20 lbs would require (most likely) a hydraulic and heavy robotic manipulator; for example, a Cincinnati Millicron robot. It can also be speculated that the motion control of a single manipulator whose hand is holding of the rocket-wrench would be very difficult.

The foregoing considerations suggest that a dual robot i.e., a manipulator system consisting of two manipulators will be well suited to the task in question [1]. Their combined weight would be about  $1/4 - 1/5$  of the total weight of a single manipulator capable of carrying the load in question. A dual robotic manipulator will be able to grasp the rocket-wrench on both sides,

and move it accurately to the desired pose. Moreover, the hands of the dual arm system will be able to hold the rocket-wrench steadily. After the rocket-wrench is at the desired pose, the bolts of the wrench assembly must be turned so as to make the jaws of the wrench tightly grasp the fuze. A dual-arm manipulator could use one arm/hand to tighten a bolt while the other arm/hand is holding the rocket-wrench assembly. Moreover, the dual arm system is versatile and adaptable to the disposal of a fuze mechanism in different types of ordnance as well.

#### **4. DUAL-ARM MANIPULATOR SYSTEM**

The dual-arm system consists of two robotic manipulators . Their bases are mounted on a platform (plate) which can be attached to an excavator as an end-effector in the same way as the bucket. The workspaces of the two robotic manipulators must partly overlap, and their intersection forms the workspace of the dual-arm system.

The dual-arm manipulator can be powered either hydraulically or electrically. The hydraulic system of the manipulator could be connected (possibly automatically) to that of the excavator. Thus, the motion of the manipulator system would be caused by means of hydraulic motors. The computer system still needs electric power. Alternatively, the entire dual-arm manipulator system and its computer will receive power from an electrical power source located on the excavator base.

Each manipulator can be chosen to have six revolute links. It should be noticed that the dual arm system holding an object that possesses zero DOF (during the transfer the object is rigid) forms a closed chain through the ground. Since the rocket-wrench is to be placed to a desired pose (position and orientation) in the 3 D-space, six variables are to be specified; thus, the task space is six-dimensional. It follows that the dual-arm system composed of two six-joint manipulators holding the rocket-wrench assembly possesses six degrees of redundancy. The redundancy of the system can be utilized advantageously, for example, to avoid obstacles.

The joints of each robotic manipulator should be equipped with encoders (or potentiometers) which will be needed to provide feedback information for the position (velocity) servoing. It would be desirable, although not necessary, to have wrist-force sensors at the wrists of the robotic manipulators. The dual-arm manipulator system must have a camera for visual feedback control. When the rocket-wrench assembly will be transferred to the desired pose, the motion for the task will be controlled either by teleoperation or autonomously. In either case, the camera images are needed to provide information about the target pose. Then, the deviation (error) of the actual pose of the rocket-wrench assembly from the desired pose can be determined. For the remote control, the camera images must be displayed to the teleoperator. For the autonomous operation, the camera images will be used to extract sufficient information for controlling the pose of the rocket-wrench assembly.

A commercial robotic manipulator which from technical viewpoint would fit to the system under consideration is PUMA 562. The load carrying capacity of a single PUMA 562 is 8 lbs (3.6

kg); thus, the dual-arm manipulator can carry about 7.2 kg. This is sufficient for the payload in the task to be performed even with two attached wrist-force sensors (if they are installed). The weight of one manipulator is 130 lbs or 58.5 kg; thus, the dual-arm manipulator system without the weight of the common platform and the computer is approximately 117 kg. The two robotic manipulators and their computer will be mounted on the common platform. Thus, the dual-arm system forms an independent system which can be controlled a teleoperator using radio communications, or it can function autonomously.

## 5. CONTROL PROBLEM IN DUAL-ARM POSITIONING SYSTEM

The dynamical model of two rigid-link manipulators holding and transferring a rigid object with zero DOF has been presented in [2,9]. The rocket-wrench assembly (the object) possesses one DOF (due to the bolts). However, during the transfer of the rocket-wrench by the dual-arm system, there is no movement of the bolts; thus, the DOF of the rocket-wrench is not utilized during the transfer, and the mathematical model of the system is the same as that of the dual-arm system transferring a rigid object. After the rocket-wrench assembly has been transferred to the desired pose, then the bolts will be rotated to tighten the assembly jaws around the fuze while the robotic system keeps the rocket-wrench assembly stationary.

For the control problem, it will be assumed that the pose of the exposed and visible fuze has been determined. The pose of the fuze can be expressed, for example, by specifying the coordinates of the center point on the top of the cylindrically shaped fuze in the world (or base) coordinate system, and the equation for the axis of the fuze. It may be noted that the axis of the fuze normally coincides with that of the ordnance when the fuze is located at the rear or front of the ordnance. In addition to the pose of the fuze, the diameter of the fuze must also be calculated from the camera images. It follows that the desired pose for the rocket-wrench can be specified in the world coordinate system.

The unit normal to the plane of the jaw assembly specifies the orientation of the rocket-wrench assembly. The center point of the jaw assembly can be used as a reference point for which the desired trajectory is planned. The desired trajectory for the motion will be determined by a planner. It consists of two straight line segments in the world (or base) coordinates. The first one starts from the initial pose  $\{p^d(0)\}$  of the reference point. The end point  $\{p^d(t_1)\}$  of this first segment is chosen on the axis of the fuze, about 10 cm outside the top plane of the cylindrically shaped fuze. The second line segment is from  $\{p^d(t_1)\}$  to the desired final point  $\{p^d(t_f)\}$ . These two line segments can then pointwise be mapped into the joint space of the manipulator using the inverse kinematic equations.

Then, the control problem is to make the reference point on the rocket-wrench assembly track the specified trajectory, or equivalently, to make the joint variables track their desired values which are determined by the inverse kinematics.

## 6. TELEOPERATED CONTROL OF DUAL-ARM MANIPULATOR SYSTEM

A typical teleoperated manipulator system consists of a human operator who controls the motion of a manipulator system on a distant site. The control signals of the operator are transmitted through a communication medium (radio link) to the remotely located manipulator system which then realizes the motion commands of the human, and performs some specified tasks.

The block diagram of a typical teleoperated manipulator system is shown in Figure 1. It is common that the velocity (position) command from the operator is fed forward to the manipulator system [11-16]. The distant manipulator system is ideally operated so that (i) it realizes the velocity commands given by the human operator, and that (ii) the human operator will be able to sense the forces at the end-effector touching the environment at the remote site (telepresence). For this purpose, the human operator must receive sensory feedback from the remotely located manipulator system. Specifically, the forces/torques acting on the end-effector (gripper) of the manipulator system are fed back to the human operator so that the telepresence can be established.

### (i) A COMMON ARCHITECTURE FOR TELEOPERATED SYSTEMS

Several architectures for the teleoperation of a single robotic manipulator have been proposed in the literature. As an example, Figure 2 shows schematically an architecture in which the velocity is transmitted from the master through a forward loop to the slave manipulator and the generalized force (force/torque) from the slave manipulator through a feedback loop to the master.

The realization of the architecture in Figure 2 for a multi-joint single manipulator without the force feedback is shown in Figure 3. It is noted that the velocity vector of the operator is converted first to the Cartesian velocity vector by means of the Jacobian matrix  $J_m$  of the master (operator). Then, the Cartesian velocity is transmitted to the remote robotic manipulator (slave), where it is then converted to the joint velocity vector by using the inverse Jacobian matrix  $J_s^{-1}$  of the slave manipulator. A computer will then control the actuators of the joints so that they will generate the desired velocity vector, which was commanded by the master.

If the system has a force feedback, the force sensed by the end-effector of the remote manipulator is first expressed in the wrist coordinate frame, and then the result is transformed to the Cartesian base coordinate system of the slave robot by means of the rotation submatrix of the homogeneous transformation matrix  $(A_6^0)^{-1}$ . The result is communicated back to the site of the human operator and converted to the joint torques by premultiplying the force vector by the transposed Jacobian matrix  $J_m^T$ ; these joint torques are then realized on the master (exoskeleton) manipulator so that they influence (usually oppose) the master's motion. Thus, the operator can "feel" the remotely sensed generalized force as if he/she were performing the task at the remote location (telepresence).

## **(ii) ARCHITECTURE FOR A TELEOPERATED DUAL-ARM SYSTEM**

The control of the gross-motion of a two-arm manipulator system can be performed in the leader-follower mode. It means that the motion of the end-effector of the leader manipulator (A) is made to track a desired trajectory specified by the master. Then the motion of the end-effector of the follower manipulator (B) is made to track the motion of manipulator (A) at a pre-determined distance and at a specified orientation. Thus, the dynamics of the dual system is not directly used in the motion control; only the dynamics of a single robotic manipulator.

An alternative approach is to consider the dual arm system holding an object as a closed-chain, and to design a controller for the overall system. Thus, the dual arm system will operate autonomously. This approach is presented in [2,9], but will not be discussed here.

## **7. CONCLUSIONS**

A dual-arm robotic system is proposed here to function as the end-effector of an excavator in order to dispose a fuze from unexploded ordnance. The specifications of the robotic manipulators are outlined.

The robotic system can be controlled by a teleoperator and/or autonomously. The task of the teleoperator is to guide the transfer of a rocket-wrench fixture to a specified pose around the fuze of the ordnance. The architecture for the teleoperation is proposed.

A robotic system which is used to remove a fuze from ordnance makes it unnecessary to have a human in the neighborhood of unexploded ordnance. In fact, the human operator in this task can stay at a remote site while the task in the neighborhood of the ordnance is being performed by the robotic system. This situation thus prevents exposing human lives to the dangers of unexploded ordnance. It represents preventive means to save lives in neutralizing unexploded ordnance.

## REFERENCES:

- [1] A. J. Koivo and G. Bekey, "Report of Workshops on Coordinated Multiple Robot Manipulators: Planning, Control and Applications," *IEEE J. on Robotics and Automation*, February 1988, pp. 91-93.  
The authors organized the foregoing NSF-sponsored workshop in San Diego, CA, January 7-9, 1987.
- [2] A. J. Koivo and M. A. Unseren, "Modeling Closed Chain Motion of Two Manipulators Holding a Rigid Object," *Mechanism and Machine Theory Journal*, November 1990, pp. 427-438.
- [3] A. J. Koivo and N. Houshangi, "Real-time Vision Servoing of a Robotic Manipulator with Self-tuning Controller," *IEEE Trans. on Systems, Man and Cybernetics*, January/February 1991, pp. 134-142
- [4] T. H. Chiu, A. J. Koivo, and R. Lewczyk, "Experiments on Manipulator Gross Motion Using Self-tuning Controller and Visual Information," *Journal of Robotic Systems*, Vol. III, No. 2, March 1986, pp. 59-70.
- [5] A. J. Koivo, "On Vision Feedback Adaptive Control of Robotic Manipulators," The 30th IEEE Conference on Decision and Control, Brighton, England, December 1991
- [6] A. J. Koivo, "Kinematics of Excavators (Backhoes) for Transferring Surface Material," *Journal of Aerospace Engineering*, Vol. 7, No. 1, January 1994, pp. 17-32.
- [7] A. J. Koivo and S. Arnautovic, "Control of Redundant Manipulators with Constraints Using a Reduced Order Model," *Automatica*, The Journal of IFAC, Vol. 30, No. 4, April 1994, pp. 665-678.
- [8] A. J. Koivo, "Planning for Automatic Excavation Operations," The 9th International Symposium on Automation and Robotics in Construction, Tokyo, Japan, June 1992.
- [9] A. J. Koivo and M. A. Unseren, "Reduced Order Model and Decoupled Control Architecture for Two Manipulators Holding a Rigid Object," *Trans. of ASME, Journal of Dynamical Systems, Measurement and Control*, Vol. 113, No. 4, December 1991, pp. 646-654.
- [10] A. J. Koivo and C.-W. Kim, "Hierarchical Classification of Surface Defects on Dusty Wood Boards," *Pattern Recognition Letters*, July 1994, pp. 713-721.
- [11] A. J. Koivo, M. C. Ramos, et. al., "Development of Motion Equations for Excavators and Backhoes," *International Journal of Intelligent Mechatronics*, 1(1) September 1994, pp. 46-55.
- [12] B. Hannaford, "A Design Framework for Teleoperators with Kinesthetic Feedback," *IEEE Trans. on Robotics and Automation*, Vol. 5, No. 4, August 1989, pp. 426-434.
- [13] Y. Strassberg, A. A. Goldenberg and J. K. Mills, "A New Control Scheme for Bilateral Teleoperating Systems: Performance Evaluation and Comparison," *Proc. of the 1992 IEEE/RSJ Intl. Conference on Intelligent Robots and Systems*, Raleigh, NC, July 1992, pp. 865-872.
- [14] S. Lee and H. S. Lee, "An Advanced Teleoperator Control System: Design and Evaluation," *Proc. of the 1992 IEEE Intl. Conference on Robotics and Automation*, Nice, France, May 1992, pp. 859-864.
- [15] R. J. Anderson and M. W. Spong, "Bilateral Control of Teleoperators with Time Delay," *IEEE Trans. on Automatic Control*, Vol. 34, No. 5, May 1989, pp. 494-501.
- [16] Y. Yokokohji and T. Yoshikawa, "Bilateral Control of Master-Slave Manipulators for Ideal Kinesthetic Coupling," *Proc. of the 1992 IEEE Intl. Conference on Robotics and Automation*, Nice, France, May 1992, pp. 849-858.
- [17] W. S. Kim, B. Hannaford, and A. K. Bejczy, "Shared Compliance Control for Time-Delayed Telemanipulation," *Proc. of the 1991 IEEE Intl. Conference on Robotics and Automation*, Cincinnati, OH, May, 1990.



Figure 1

A teleoperated master/slave  
manipulator system

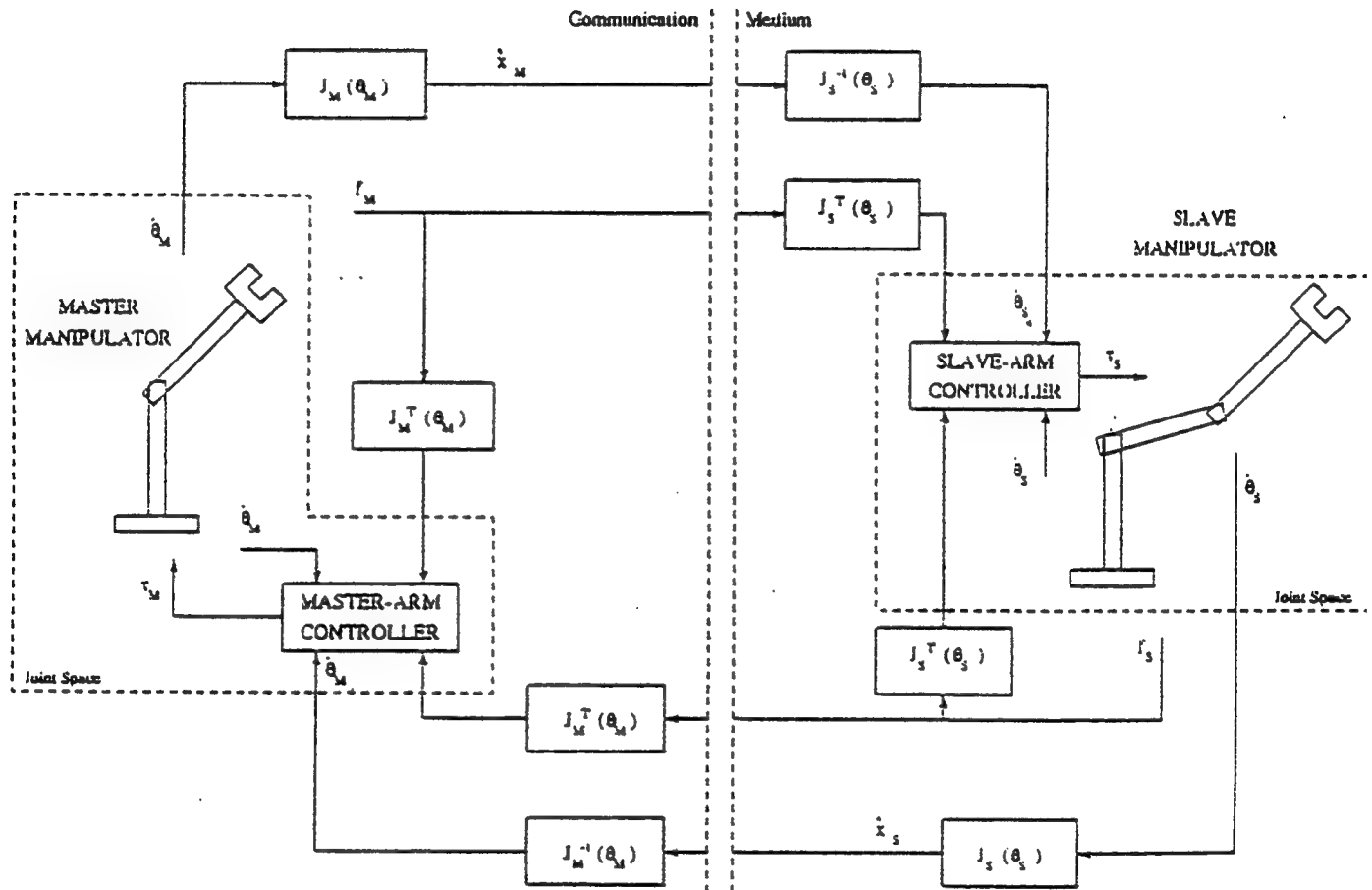


Figure 2 A Multi-dimensional Teleoperator System

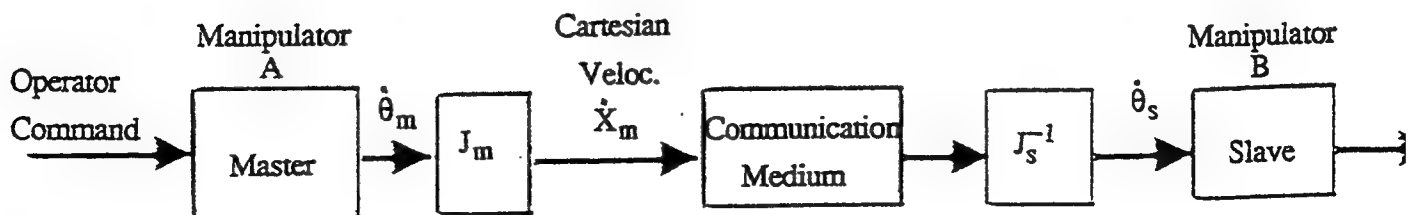


Figure 3

Teleoperated master/slave system  
without force feedback



## PART II: ADAPTIVE SELF-TUNING CONTROL OF EXCAVATORS

A. J. Koivo and Al Nease

### 1. INTRODUCTION

The semi-autonomous or automatic computer control is essential to improve the productivity and the effective utilization of expensive construction robots [1,2]. The machines must often perform in hostile and/or unfavorable working conditions, for example, in environment polluted by unexploded ordnance. In order to design automatic control schemes for the operations of the construction robots, the kinematics and dynamics of the machine motions must be well understood. Recent developments of the kinematic relations and dynamic models of the construction manipulators, specifically those of the excavators, [1,2,3] have established the foundation for a systematic controller design. The work reported here presents the design of adaptive self-tuning controller for the gross-motion of robotic excavators performing tasks automatically. The control scheme allows also human interactions in controlling the motion.

The usual task of an excavator is to free and/or remove surface material (e.g., soil, coal) from its original location and to transfer it to another location by pushing, pulling, and/or lifting it in a bucket. The execution of this task is usually controlled by a human operator who guides the motion of the machine manually by using the visual feedback provided through his/her own eyes. In many current applications of excavators, the semi-autonomous or automatic operation of the machine is desirable and sometimes even necessary; for example, in the transfer of soil from the neighborhood surrounding unexploded ordnance, or mining products from underground sites, or in the removal of poisonous or radioactive wastes. In the semi-autonomous operation, a human tele-operator may guide the motion of the machine over a certain time period so as to make it perform parts of the task and then a computer may control the machine over the other times of the motion so as to execute the task automatically. In the automatic computer controlled motion, the bucket of the excavator tracks a desired trajectory specifying the position and the digging angle, i.e., the pose of the bucket, which corresponds to specific values of the angular positions of the joint shafts in the excavator. The values of these joint variables, in turn, are determined by the lengths of the pistons in the hydraulic actuators. The mathematical relations between these variables are described by the kinematic relations of the machine [1].

The previously reported studies on excavators are mainly qualitative while technical aspects are usually not described. An entire system for the automatic or semi-automatic operation of an excavator is presented in [4]. This description includes a hydraulically driven excavator, teleoperation, computer control, which is supplemented with manual override capability, position and force feedback control. Kinematic relations between three joint angles and the position of the bucket are presented in [4] using the geometric configuration of the excavator arm in a fixed (world) coordinate system. However, no other details on the system are given. The use of a position and sensor (force and vision) feedback is discussed qualitatively in [5] without presenting technical details. Similarly, a vision feedback system for an excavator

designed for rapid runway repairs is presented in [5] by qualitatively characterizing the system components. This system has successfully been tested in practice, but no technical details are given in [5]. Although the forces between the soil and a tool (bucket) during the digging operations are studied in [6,7], the kinematic relations between the bucket pose, the joint variables, and the lengths of the pistons of the actuators are not described in [6,7]. The first systematic and complete approach to describe the kinematics of the excavators using the Denavit-Hartenberg (D-H) guidelines is presented in [1].

The development of equations of motion for excavators has not attracted attention of researchers until recently. The dynamic modeling of the excavator motion is presented in [2]. The equations of motion are described in the standard form commonly encountered in the robotics literature. They will be used here in the simulation and testing of the designed self-tuning controller.

## 2. DYNAMIC MODEL FOR EXCAVATOR

The equations of motion for an excavator can be written as Newton-Euler (N-E) equations [2]. After some manipulations, they can be expressed as a set of second-order nonlinear differential equations in the following form:

$$D(\theta)\ddot{\theta} + C(\theta, \dot{\theta})\dot{\theta} + G(\theta) + B(\dot{\theta}) = \Gamma(\theta)F - F_{\text{load}}(F_t, F_n) \quad (1)$$

where  $\theta = [\theta_1 \ \theta_2 \ \theta_3 \ \theta_4]^T$  and  $\theta_i, i = 1, 2, 3, 4$  represents the shaft angle of joint  $i$ , and the superscript  $T$  refers to the transposition. The  $(4 \times 4)$  matrix  $\Gamma(\theta)$  is a function of the moment arms; vector  $F = [F_b \ F_{BE} \ F_{FI} \ F_{JK}]^T$  specifies the forces on the pistons of the hydraulic actuators which produce the torques acting on the joint shafts; the first component  $F_b \equiv 0$  since the first joint is not moved. Term  $F_{\text{load}}$  is determined by the external forces  $F_n$  and  $F_t$  acting on the bucket due to soil. Vector  $G(\theta)$  describes the gravitational torques,  $C(\theta, \dot{\theta})\dot{\theta}$  is determined by the Coriolis and centripetal effects,  $D(\theta)$  is the (pseudo) inertial matrix.

Specifically,  $G(\theta)$ ,  $C(\theta, \dot{\theta})$  and  $D(\theta)$  are given by the following expressions:

$$G(\theta) = [G_1 \ G_2 \ G_3 \ G_4]^T \quad (2)$$

$$C(\theta, \dot{\theta})\dot{\theta} = \begin{bmatrix} C_{11} & C_{12} & C_{13} & C_{14} \\ C_{21} & C_{22} & C_{23} & C_{24} \\ C_{31} & C_{32} & C_{33} & C_{34} \\ C_{41} & C_{42} & C_{43} & C_{44} \end{bmatrix} \begin{bmatrix} \dot{\theta}_1 \\ \dot{\theta}_2 \\ \dot{\theta}_3 \\ \dot{\theta}_4 \end{bmatrix} \quad (3)$$

where in equation (2)  $G_2 = -m_4 g [l_2 c_2 + l_3 c_{23} + L_{03} G_4 \cos(\theta_{234} + \sigma_4)] - m_3 g [l_2 c_2 + L_{02} G_3 \cos(\theta_{23} + \sigma_5)] - m_2 g L_{01} G_2 \cos(\theta_2 + \sigma_9)$ ;  $G_3 = -m_4 g [l_3 c_{23} + L_{03} G_4 \cos(\theta_{234} + \sigma_4)] - m_3 g L_{02} G_3 \cos(\theta_{23} + \sigma_5)$ ;  $G_4 = -m_4 g L_{03} G_4 \cos(\theta_{234} + \sigma_4)$ ; and in equation (3),  $C_{22} = -2[d' +$

$k\sin(\theta_{34} + \sigma_4)]\dot{\theta}_3 - 2[k\sin(\theta_{34} + \sigma_4) + n\sin(\theta_4 + \sigma_4)]\dot{\theta}_4$ , and  $d' = m_3 l_2 L_{02G3} \sin(\theta_3 + \sigma_5) + m_4 l_2 l_3 S_3$ ,  $k = m_4 l_2 L_{03G4}$ ,  $n = m_4 l_3 L_{03G4}$ ;  $C_{23} = -[d' + k\sin(\theta_{34} + \sigma_4)]\dot{\theta}_3 - 2[k\sin(\theta_{34} + \sigma_4) + n\sin(\theta_4 + \sigma_4)]\dot{\theta}_4$ ;  $C_{24} = -[k\sin(\theta_{34} + \sigma_4) + n\sin(\theta_4 + \sigma_4)]\dot{\theta}_4$ ;  $C_{32} = [d' + k\sin(\theta_{34} + \sigma_4)]\dot{\theta}_2 - [n\sin(\theta_4 + \sigma_4)]\dot{\theta}_4$ ;  $C_{33} = -[n\sin(\theta_4 + \sigma_4)]\dot{\theta}_4$ ;  $C_{34} = -[n\sin(\theta_4 + \sigma_4)](\dot{\theta}_2 + \dot{\theta}_3 + \dot{\theta}_4)$ ;  $C_{42} = [k\sin(\theta_{34} + \sigma_4) + n\sin(\theta_4 + \sigma_4)]\dot{\theta}_2 - [n\sin(\theta_4 + \sigma_4)]\dot{\theta}_3$ ;  $C_{43} = [n\sin(\theta_4 + \sigma_4)](\dot{\theta}_2 + \dot{\theta}_3)$ ;  $C_{44} = 0$ . Moreover,

$$D(\theta) = \begin{bmatrix} D_{11} & D_{12} & D_{13} & D_{14} \\ D_{21} & D_{22} & D_{23} & D_{24} \\ D_{31} & D_{32} & D_{33} & D_{34} \\ D_{41} & D_{42} & D_{43} & c \end{bmatrix} \quad (4)$$

where  $D_{22} = \bar{a}_1 + 2d + 2n\cos(\theta_4 + \sigma_4) + 2k\cos(\theta_{34} + \sigma_4)$ ;  $D_{23} = D_{32} = \bar{a}_2 + d + 2n\cos(\theta_4 + \sigma_4) + k\cos(\theta_{34} + \sigma_4)$ ;  $D_{24} = D_{42} = c + n\cos(\theta_4 + \sigma_4) + k\cos(\theta_{34} + \sigma_4)$ ;  $D_{33} = \bar{a}_2 + 2n\cos(\theta_4 + \sigma_4)$ ;  $D_{34} = D_{43} = c + n\cos(\theta_4 + \sigma_4)$ ;  $\bar{a}_1 = a + b + c$ ;  $\bar{a}_2 = b + c$ ;  $a = m_2 L_{01G2}^2 + I_{02} + (m_3 + m_4)l_2^2$ ;  $b = m_3 L_{02G3}^2 + I_{03} + m_4 l_3^2$ ;  $c = m_4 L_{03G4}^2 + I_{04}$ ;  $d = m_3 l_2 L_{02G3} \cos(\theta_3 + \sigma_5) + m_4 l_2 l_3 c_3$ .  $m_i$  and  $l_i$  are the mass and length of link  $i$ , respectively, and  $I_i$  is its second-order inertial moment about the rotational axis through the gravity center;  $g = -9.81 \text{ m/s}^2$ ;  $c_i = \cos \theta_i$ ;  $c_{ij} = \cos(\theta_i + \theta_j)$ ;  $c_{ijk} = \cos(\theta_i + \theta_j + \theta_k)$  and  $s_i, s_{ij}, s_{ijk}$  are the corresponding expressions of the sin-function. Moreover,

$$\Gamma(\theta) = \begin{bmatrix} \Gamma_{11} & \Gamma_{12} & \Gamma_{13} & \Gamma_{14} \\ 0 & \Gamma_{22} & \Gamma_{23} & \Gamma_{24} \\ 0 & 0 & \Gamma_{33} & \Gamma_{34} \\ 0 & 0 & 0 & \Gamma_{44} \end{bmatrix} \quad (5)$$

where the elements of  $\Gamma(\theta)$  are defined as

$$\begin{aligned} \Gamma_{22} &= L_{01B} \sin(\rho - \theta_2 - \sigma_{11}); \Gamma_{23} = L_{01I} \sin(\theta_3 + \sigma_{10} - \gamma_2) - l_2 \sin(\theta_3 - \gamma_2) \\ \Gamma_{24} &= L_{0102} \sin(\gamma_1 + \theta_3) - L_{0102} \sin(\varepsilon_4 + \theta_3) \left[ \frac{\sin \gamma_1 - \cos \gamma_1 \tan \varepsilon_4}{\sin(\theta_{23} - \varepsilon_5) + \tan \varepsilon_4 \cos(\theta_{23} - \varepsilon_5)} \right] \\ &\quad - L_{0102} \sin(\gamma_2 - \varepsilon_5) \left[ \frac{\cos \gamma_1 \tan(\theta_{23} - \varepsilon_5) + \sin \gamma_1}{\cos \varepsilon_4 \tan(\theta_{23} - \varepsilon_5) + \sin \varepsilon_4} \right] \\ \Gamma_{33} &= L_{02F} \sin(\sigma_8 - \gamma_2) \end{aligned}$$

$$\begin{aligned}
\Gamma_{34} = & -L_{02}L\sin(\epsilon_4 - \sigma_6) \left[ \frac{\tan(\theta_{23} - \epsilon_5)\cos\gamma_1 + \sin\gamma_1}{\sin\epsilon_4 + \cos\epsilon_4 \tan(\theta_{23} - \epsilon_5)} \right] + L_{02}J\sin(\sigma_7 - \gamma_1) \\
& + l_3 \sin(\epsilon_5 - \theta_{23}) \left[ \frac{\tan\epsilon_4 \cos\gamma_1 - \sin\gamma_1}{\sin(\theta_{23} - \epsilon_5) - \tan\epsilon_4 \cos(\theta_{23} - \epsilon_5)} \right] \\
\Gamma_{44} = & L_{03}P\sin[\epsilon_5 - (\underbrace{PO_3O_4}_{\theta_{234}} - \theta_{234})] \left[ \frac{\tan\epsilon_4 \cos\gamma_1 - \sin\gamma_1}{\sin(\theta_{23} - \epsilon_5) - \tan\epsilon_4 \cos(\theta_{23} - \epsilon_5)} \right] \\
F_{load}(F_t, F_n) = & J^T(\theta)F_{Lo} \tag{6}
\end{aligned}$$

where  $F_{Lo} = [F_{to} \ F_{no}]^T$ ,  $F_{Lo} = (A_1^4)_R [F_t \ F_n]^T$ , and  $(A_1^4)_R$  is the rotation submatrix of the homogeneous transformation matrix  $A_1^4$ . Moreover,  $J^T(\theta)$ , the transpose of the Jacobian matrix of the excavator that relates the forces/torques acting on the bucket to the joint torques is specified as:

$$J(\theta) = \begin{bmatrix} -a_4 s_{234} - a_3 s_{23} - a_2 s_{23} & -a_4 s_{234} - a_3 s_{23} & -a_4 s_{234} \\ a_4 c_{234} + a_3 c_{23} + a_2 c_2 & a_4 c_{234} + a_3 c_{23} & a_4 c_{234} \\ 1 & 1 & 1 \end{bmatrix} \tag{7}$$

and  $a_i$  is the perpendicular distance between the  $Z_{i-1}$ - and  $Z_i$ - axes in the chosen coordinate frame system. Equivalently, one may write  $F_{Lo}$  as

$$F_{Lo} = \begin{bmatrix} \cos\theta_{dg} & -\sin\theta_{dg} \\ \sin\theta_{dg} & \cos\theta_{dg} \end{bmatrix} \begin{bmatrix} F_t \\ F_n \end{bmatrix} \tag{8}$$

In equations (1) through (5), elements  $D_{1i}, D_{i1}, C_{1i}, C_{i1}, \Gamma_{1i}$ ,  $i = 1, 2, 3, 4$   $G_1$  and  $F_1$  are not specified here because the joint variable  $\theta_1$  is not changed during the digging operation.

One may observe that the inertia matrix  $D(\theta)$  is symmetric as in the model of a robotic manipulator and positive definite (due to the kinetic energy expression for the excavator motion).

### 3. PROBLEM STATEMENT FOR SELF-TUNING CONTROLLER DESIGN

The motion of the joints of an excavator is governed by equation (1). It is assumed that the planner of the excavator motion has calculated the desired trajectories for all joints, i.e., the trajectory  $q_i^d(t)$  for joint  $i$ ,  $i = 2, 3, 4$  over the duration  $0 \leq t \leq t_f$  of the motion has been determined. It is assumed that the actual positions of the joint shafts can be measured, for example, by means of encoders or potentiometers.

Then the problem is to design a feedback controller such that the motion  $q_i(t)$  of each joint  $i = 2, 3, 4$  in the excavator will track the specified trajectory  $q_i^d(t)$ .

#### 4. ADAPTIVE SELF-TUNING CONTROLLER DESIGN

Since the controller to be designed will be implemented on a digital computer, and the encoder reading are discrete in-time, the design will be performed using a discrete-time approach. If the model in equation (1) were discretized in-time using first-order (forward) difference approximations for the derivatives, the resulting model would be a set of second-order difference equations. Therefore, the following discrete time-series equation will be assumed to model the input-output data (input = joint torque, output = joint position) of joint  $i$ ,  $i = 2,3,4$ :

$$y_i(k+2) = a_i^0 + a_i^1 y_i(k+1) + a_i^2 y_i(k) + b_i^1 u_i(k+1) + e_i(k+2) \quad (9)$$

where  $k = 0,1,2,\dots$ ,  $y_i(k)$  represents the angular position of joint  $i$  of the excavator at time  $kT$ ,  $u_i(k)$  signifies the input torque at joint  $i$  at time  $kT$ ,  $b_i^1$  and  $a_i^n$ ,  $n = 0,1,2$  are unknown parameters in the dynamic model of joint  $i$ , and they are to be estimated on the basis of the available measurements. The random variable  $e_i(k+2)$  is Gaussian, has zero mean and finite variance; and it signifies modelling errors. The sampling period  $T$  is not shown in the arguments of the variables in equation (9) for convenience.

It may be noticed that the assumed model is linear and does not contain terms which would represent coupling effects between adjacent joints. Thus, the interactions between the joints during the motion are neglected. This situation is called in the robotics literature independent joint dynamics (control). If it is necessary, coupling terms can be included in equation (9), and the following approach can still be used without any essential modifications.

The controller will be designed so that the shaft position  $q_i(k)$  of joint  $i$  of the excavator tracks the desired trajectory  $q_i^d(k)$ . It can be achieved if the following criterion is minimized:

$$C_i[u_i(k)] = E \left\{ [y_i(k+2) - y_i^d(k+2)]^2 / 2 + w_i u_i^2(k) / 2 \mid \Sigma \right\} \quad (10)$$

where  $E \left\{ \bullet \mid \Sigma \right\}$  signifies the expectation operation conditioned on the available measurements at time  $kT$ , and  $w_i$  specifies the weighting factor of  $u_i^2(k)$  (energy term).

The problem is now to minimize  $C_i[u_i(k)]$  with respect  $u_i(k)$  while satisfying the model constraint, equation (9).

Since the parameters in equation (9) are not known precisely, they will be estimated on the basis of the measurement values, for example, by minimizing the standard least squared error criterion. Then the estimates of the parameters will be used in the constraint equation when criterion  $C_i[u_i(k)]$  is minimized.

To describe the recursive parameter estimator equations, equation (9) is rewritten as

$$y_i(k+2) = [y_i(k+1) \ y_i(k) \ u_i(k+1) \ 1] \alpha + e_i(k+2) \quad (11a)$$

when  $\alpha_i = [\alpha_i^1 \ \alpha_i^2 \ b_i^1 \ 1]^T$ , or simply with obvious definitions of the variables as

$$y_i(k+2) = \phi_i^T(k+1) \alpha_i + e_i(k+2) \quad (11b)$$

The well-known recursive parameters estimation equations can then be written as

$$\hat{\alpha}_i(k) = \hat{\alpha}_i(k-1) + P_i(k) \phi_i(k-1) [y_i(k) - \phi_i^T(k-1) \hat{\alpha}_i(k-1)] \quad (12)$$

$$P_i(k) = \frac{1}{\gamma_i} \left[ P_i(k-1) - \frac{P_i(k-1) \phi_i(k-1) \phi_i^T(k-1) P_i(k-1)}{\gamma_i + \phi_i^T(k-1) P_i(k-1) \phi_i(k-1)} \right] \quad (13)$$

where  $\gamma_i$  represents a forgetting factor, matrix  $P_i(k) = P_i^T(k)$  is the covariance of the estimation error, and  $\hat{\alpha}_i(k)$  signifies the estimate of parameter  $\alpha$  at time  $kT$ . Equations (12) and (13) can be solved on-line.

The calculated estimate  $\hat{\alpha}_i(k+1)$  is then substituted into the model, equation (11a). The resulting model with these parameter values is used while criterion  $C_i[u_i(k)]$  in equation (10) is minimized. The minimization yields the following control law:

$$u_i(k) = -b_i^1 [\hat{a}_i^0 + \hat{a}_i^1 y_i(k) + \hat{a}_i^2 y_i(k-1) - y_i^d(k+1)] / [(\hat{b}_i^1)^2 + w] \quad (14)$$

Equation (10) specifies the adaptive self-tuning controller. Since all terms are known at time  $kT$ , the controller is realizable.

## 5. SIMULATION RESULTS

The motion of the excavator was simulated on the basis of equation (1). The reaction force  $F_r$  acting on the edge of the bucket is calculated as explained in Appendix A [7]; it follows then that  $F_t = F_r \cos(0.1 \text{ rad})$  and  $F_n = -F_r \sin(0.1 \text{ rad})$ . The numerical values of the parameters for simulating the dynamical model in equation (1) are given in Appendix A.

The time-series model chosen was

$$y_i(k+2) = a_i^0 + a_i^1 y_i(k+1) + a_i^2 y_i(k) + b_i u_i(k+1) + e_i(k+2) \quad (15)$$

where  $y_i(k)$  represents the shaft position of joint  $i$ ,  $i = 2, 3, 4$  at time  $kT$ , coefficients  $a_i^0$ ,  $a_i^1$  and  $a_i^2$  are parameters to be estimated by the least squared error method, and  $e_i(k+2)$  signifies equation (modeling) error which is assumed to be white zero-mean Gaussian variable with finite variance. Thus, independent joint dynamics are assumed.

The tracking of the specified trajectory  $y_i^d(k)$  calculated by the planner on the basis of the inverse kinematic equations is achieved by minimizing

$$C_i[u_i(k)] = E \left\{ [y_i(k+2) + y_i^d(k+2)]^2 / 2 + w_i u_i^2(k) / 2 \mid \Sigma \right\} \quad (16)$$

where the weighting factor  $w_i$  is chosen as 0.08 after some experimentation. The same value was used for  $i = 2, 3, 4$  in the design.

The unknown parameters  $a_i^n$ ,  $n = 0, 1, 2$  and  $b_i$  are first estimated using the recursive equations (12) and (13). These calculated values  $\hat{a}_i^n$  and  $\hat{b}_i$  are substituted into equation (11a). Then the optimal self-tuning controller is determined by minimizing  $C_i[u_i(k)]$  while satisfying the time-series model with the parameter estimates approximating the true parameters. The minimizing controller is specified as

$$u_i(k) = [ - \hat{b}_i / (\hat{b}_i^2 + w_i) ] [ \hat{a}_i^0 + \hat{a}_i^1 y_i(k+1) + \hat{a}_i^2 y_i(k) - y_i^d(k+2) ] \quad (17)$$

where the last available estimates are used for the parameters. Thus, the self-tuning controller can be implemented by constructing a feedback loop on the basis of equation (14). The control architecture is displayed in Figure 1.

The motion of the excavator was simulated using equation (1). The self-tuning controller in equation (17) was employed as shown in Figure 1. The simulation results are displayed in Figure 2. After an initial learning period, the tracking of the desired trajectory is very good. When a disturbance pulse in the middle of the motion duration was introduced, which simulates a sudden increase in the reaction force acting on the bucket, the adaptive self-tuning controller is able to overcome the disturbance effects in a short time, as seen in figure 4.

## 6. CONCLUSIONS

To design a self-tuning controller, the input-output relation of an excavator is described as a time-series model of ARX-type (autoregressive model with external input). The parameters of the model are estimated by the least squared error method. The controller is determined by minimizing the squared tracking error and the energy used during the motion. The controller is in a feedback form, and it possesses adaptive property by adjusting the gains on the basis of the last parameter estimates. Simulations are presented to demonstrate the performance of the self-tuning controller.

## APPENDIX A: Reaction Force and Numerical Values for Model Parameters.

(i) Reaction Force: During the digging operation, the reaction force on the edge of the bucket is determined by [8] as follows:

$$F_r = k_p [k_s b h + \mu N + \epsilon (1 + \frac{V_s}{V_b}) b h \sum_i \Delta x_i] \quad (A.1)$$

where  $k_p$  and  $k_s$  are specific resistances in cutting silty clay; constants  $b$  and  $h$  are the width and thickness of the cut slice of soil, respectively;  $\mu$  is the coefficient of friction between the bucket and the soil;  $N$  is the pressure force of the bucket with the soil;  $\epsilon$  is the coefficient of resistance experienced in filling the bucket during the movement of the prism of soil;  $V_s$  and  $V_b$  are the volumes of the prism of soil and the bucket, respectively; and  $\Delta x$  is the increment along the horizontal axis (in meters).

The reaction force is defined to be parallel to the digging direction. Its horizontal ( $F_h$ ) and vertical ( $F_v$ ) components with respect to the soil are:  $F_h = F_r \cos(\theta_{dg} - 0.1)$ ;  $F_v = F_r \sin(\theta_{dg} - 0.1)$  (Vähä et al. 1991). Then the tangential component  $F_t = F_r \cos(0.1)$  and the normal component  $F_n = -F_r \sin(0.1)$  can be calculated.

(ii) Numerical Values of Model Parameters: The numerical values of the parameters used in the simulation are: The lengths of the links:  $a_1 = 0.05$  m,  $a_2 = 5.16$  m,  $a_3 = 2.59$  m,  $a_4 = 1.33$  m. The controller gains:  $K_p = 638863$  and  $K_v = 800$ . The distances between points described by the subscripts are:  $d_{AB} = 2.71$  m,  $d_{AH} = 0.56$ ,  $d_{AI} = 2.6$  m,  $d_{AP} = 2.5$  m,  $d_{CF} = 0.77$  m,  $d_{CI} = 2.8$  m,  $d_{CJ} = 0.63$  m,  $d_{CL} = 0.63$  m,  $d_{CQ} = 0.37$  m,  $d_{DG} = 0.4$  m,  $d_{DP} = 0.5$  m,  $d_{DR} = 0.65$  m,  $d_{DJ} = 2.23$  m,  $d_{EH} = 0.42$  m.

The angles between the line indicated by the two subscripted letters are: DR-DN:  $\sigma_4 = 0.3933$  rad; DQ-QC:  $\sigma_5 = 0.3316$  rad; LC-CD:  $\sigma_6 = 0.1536$  rad; JC-CD:  $\sigma_7 = 1.4661$  rad; FC-CD:  $\sigma_8 = 2.7105$  rad; CA-AI:  $\sigma_{10} = 0.4782$  rad; BA-AC:  $\sigma_{11} = 0.4957$  rad; DJ-JK:  $\gamma_1 = \pi/6 - \theta_4$  rad; DF-FI:  $\gamma_2 = \pi - \theta_3$  rad;  $\epsilon_2 = \theta_4$ ; DL-LK:  $\epsilon_4 = \pi/2 - \theta_4$  rad; DP-PK:  $\epsilon_5 = \pi/6 - \theta_4 / 2$  rad; BE-EH:  $\pi - \rho = \tan^{-1} [(d_{AH} + d_{AB} \sin(\theta_2)) / (d_{EH} - d_{AB} \cos(\theta_2))]$ . The inertial moments:  $I_{02} = 14250.6 \text{ kgm}^2$ ;  $I_{03} = 727.7 \text{ kg m}^2$ ;  $I_{04} = 224.6 \text{ kg m}^2$ . The link masses:  $m_2 = 1566$  kg,  $m_3 = 735$  kg,  $m_4 = 432$  kg. The specific resistances to cutting for silty clay:  $k_p = 1.0005$ ,  $k_s = 5500$ . The width and thickness of the cut slice of soil respectively:  $b = 0.61$  m,  $h = 0.5$  m. The coefficient of friction of the bucket with the soil:  $\mu = 0.01$ . The pressure force of the bucket with the soil:  $N = 1 \text{ kgm/s}^2$ . The coefficient of resistance to fill the bucket and movement of the prism of soil:  $\epsilon = 55000 \text{ kg/(m}^2\text{s}^2)$ . The volume of the bucket:  $V_b = 0.58 \text{ m}^3$ . The soil density:  $\gamma_{\text{siltyclay}} = 1921.8 \text{ kg/m}^3$ .



## REFERENCES

- [1] A. J. Koivo, "Kinematics of Excavators (Backhoes) for Transferring Surface Material," *Journal of Aerospace Engineering*, Vol. 7, No. 1, January 1994, pp. 17-32.
- [2] A. J. Koivo, M. C. Ramos, et. al., "Development of Motion Equations for Excavators and Backhoes," *International Journal of Intelligent Mechatronics*, 1(1) September 1994, pp. 46-55.
- [3] A. J. Koivo, "Planning for Automatic Excavation Operations," The 9th International Symposium on Automation and Robotics in Construction, Tokyo, Japan, June 1992.
- [4] D. W. Seward, D. A. Bradley and R. H. Bracewell (1988). "The Development of Research Models for Automatic Excavation," *Proc. 5th Intl. Symp. on Robotics in Constr.*, Intl. Assoc. for Automation of Robotics in Construction, pp. 703-708.
- [5] W. P. Wohlford, F. D. Griswold and B. D. Bode (1990). "New Capability for Remote Controlled Excavation," *Proc. 38th Conf. on Remote Systems Tech.*, American Nuclear Society, pp. 228-232.
- [6] L. E. Bernold (1991). "Experimental Studies on Mechanics of Lunar Excavation," *J. of Aerosp. Engrng., ASCE*, 4(1), pp. 9-22.
- [7] M. D. Bullock, M. S. Apte and I. J. Oppenheim (1990). "Force and Geometry Constraints in Robot Excavation," *Proc. Space '90, Engrng., Constr. and Operations in Space II*, ASCE, pp. 960-969.
- [8] T. V. Alekseeva, K. A. Artem'ev, A. A. Bromberg, R. I. Voitsekhouskii and N. A. Ul'yanov (1985). *Machines for Earthmoving Work, Theory and Calculations*, Amerind Publishing Co. Pvt. Ltd., New Delhi, India.

**Kinetics of Network Reformation in Hydrolytic Degraded AFR700B Polyimide Resin**

**Andre Y. Lee  
Assistant Professor  
Materials Science and Mechanics Department**

**Michigan State University  
Lansing, MI**

**Final Report for:  
Summer Faculty Research Program  
Wright Laboratory**

**Sponsored by:  
Air Force Office of Scientific Research  
Bolling AFB, Washington, DC**

**and**

**Wright Laboratory**

**August 1995**

# **KINETICS OF NETWORK REFORMATION IN HYDROLYTIC DEGRADED AFR700B POLYIMIDE RESIN**

Andre Lee  
Assistant Professor  
Department of Materials Science and mechanics  
Michigan state University

## **Abstract**

The kinetics of high temperature curing of hydrolytic degraded AFR700B resin was investigated using dynamic mechanical spectrometry. Upon exposure to moisture, significant degradation on physical and mechanical properties of newly developed high temperature PMR type polyimide resins - AFR700B was observed. In this study we are particular interested in the possibility to re-postcuring of these hydrolyzed resins. Viscoelastic experiments were performed at temperature of 400 °C and show the isochronal elastic modulus increase at curing time increase. Furthermore, it was possible to perform a time-curing time superposition of these isothermal elastic modulus curves. This time-curing time shift rate can be related to the kinetics of polyimide network formation. Interestingly, in this study we found that the time-curing time shift rate for the hydrolytic degraded samples was identical to the time-curing time shift factor of freshly formulated samples post-cured at the same temperature. This observation does not require the chemistry of cross linking to be the same but only they are of the same control mechanism - diffusion control.

# KINETICS OF NETWORK REFORMATION IN HYDROLYTIC DEGRADED AFR700B POLYIMIDE RESIN

Andre Lee

## INTRODUCTION

In recent years, there has been significant development in imide polymers as the state-of-the-art non-metallic high performance structural materials for aerospace applications. Based on the in-situ Polymerization of Monomer Reactants (PMR) technique, imide polymers can be formulated to exhibit high glass transition, i.e.  $T_g > 400^\circ\text{C}$  as determined using dynamical mechanical method and differential scanning calorimetry, and easy processibility as the matrix in carbon fiber reinforced composites.

As a structural material for some of aerospace applications, the long-term stability of structural properties are of considerable importance. Since some structural properties depend on the physical characteristic of the polymer matrix, we are concern about the long term exposure to heat, fuels, and solvents or other chemicals that degrade the material properties. For some cases, the change in physical and mechanical properties at a particular operating environment are predictable and is accounted in the design allowable. Most of all, these predictable changes are reversible and do not alter any molecular or chemical state of the polymer matrix. However, there are other types of degradation due to the environmental exposure which are irreversible where the molecular and/or chemical state of the polymer is altered. For these types of degradation, the development of detail property-environmental exposure correlation is not economic feasible since we need to know detail history of environmental exposure and may not be a useful design methodology. Rather, a fundamental understanding on the impact of environmental exposure to the rate of property degradation is required. For carbon fiber reinforced polyimides matrix composites, the long term exposure to moisture is such a situation. Specifically, molecular and chemical degradation of imide polymers due to hydrolysis is "irreversible" under the normal service

conditions. As portion of cross linked network being destroy, which cause a significant decreases in the glass transition, i.e. as large as 80 °C change has been reported, and this potentially leading to catastrophic failure. As mentioned above, imides polymers are been considered as matrix resins for high temperature aerospace applications where property stability is required, a detail understanding of network degradation as effects of environmental exposure is of considerable importance.

As part of larger study for fundamental understanding of degradation of polyimides by hydrolysis, the objective of this study was focus on the change in the viscoelastic properties of an advanced high temperature polyimide - AFR700B. In particular we are interested in:

- (1) Measure the rate of change in viscoelastic properties of AFR700B during the post-curing stage. It is believe the chemical cross linking of AFR700B occurs at this stage, where AFR700B polymer changes from a entangled polymer melt (thermoplastic) to a crosslinked thermoset.
- (2) Determine the impact of hydrolytic degradation on physical and molecular properties of AFR700B resin.
- (3) Study the re-curing process of degraded AFR700B and compare the rate of change in viscoelastic properties to the post-curing process.

## **BACKGROUND**

### **Materials**

The synthesis of AFR700B is based on a technique known as in-situ Polymerization of Monomer Reactants (PMR) developed in the early 70's. The PMR process comprises of dissolving a monoalkyl ester of 5-norbornene-2,3-di-carboxylic acid (nadid ester, NE), an aromatic diamine, and a dialkyl ester of an aromatic tetracarboxylic acid in a low-boiling alkyl alcohol, e.g. methanol or ethanol. In the PMR approach, the aromatic diamine is reacted to di-ester and form an oligomeric polyamic acid that undergoes

cyclocondensation reaction at an elevated temperature followed by a chain extension and cross linking reaction at an even higher temperature. This chemistry retains solubility while circumventing the condensation product transport out of the high molecular weight cured polymer. It is important that the end-group can not react significantly until the cyclocondensation reaction is complete, otherwise the condensation products can not diffuse out of the system. The high solubility of monomeric solution is used to impregnate the reinforcing carbon/glass fibers; in-situ polymerization through the nadic end-group occurs directly on the fiber surface, producing a fiber reinforced composite material with desired thermal and mechanical properties. The attractive features of polymers produced through PMR technique include (a) the use of low molecular weight, low viscosity monomers where complete wetting of fibers is possible, (b) the use of a low-boiling solvent which is completely removed during polymerization process, (c) little or no evolution of volatile materials during the final curing step.

The most well known of PMR polyimide is the PMR-15. However, PMR-15 has several deficiencies which include inadequate resin flow in the fabrication of thick and complicated structures, microcracking, health and safety concerns of 4,4' - methylene dianiline (MDA), and lastly, lack of thermo-oxidative stability at 370 °C. Air Force study have shown that PMR type resins based on a hexafluoroisopropylidene (6-FDE) and p-PDA diamine offer balance of thermo-oxidative stability, high glass transition, and processability. There are several commercial variants of 6-FDE PMR resins, namely NASA PMR-II-30, PMR-II-50, V-CAP-50, V-CAP-70, DuPont Avimid-N, AFR700B, and AFR700B. In this study AFR700B is used.

#### Viscoelasticity

Polymeric materials exhibit mechanical response characteristics which are outside of the scope of such theories of mechanical behavior as elasticity and viscosity. To be more specific, the theory of elasticity many account for materials which have a capacity to store

mechanical energy with no dissipation of the energy. On the other hand, a Newtonian viscous fluid in a nonhydrostatic stress state implies a capacity for dissipating energy, but none for storing it. But, then, materials which must be outside the scope of these two theories are those for which some, but not all, of the work done to deform them, can be recovered. Such materials possess a capacity to both store and dissipate mechanical energy.

For a viscoelastic material subjected to steady state harmonic oscillatory strain,  $\gamma(\omega)$ , and if the viscoelastic behavior is linear, it is found the stress response,  $\sigma(\omega)$ , will alternate sinusoidally with same frequency but will be out of phase with the apply strain. Furthermore, the dynamic modulus,  $G^*(\omega) = \sigma(\omega) / \gamma(\omega)$ , can be separated into two frequency-dependent function - the storage modulus,  $G'(\omega)$ , and the loss modulus  $G''(\omega)$ .

The modulus  $G'(\omega)$  is defined as the stress in phase with the sinusoidal apply strain divided by the strain. It is a measure of the energy stored and recovered per cycle, when different systems are compared at the same strain amplitude. The modulus  $G''(\omega)$  is the stress 90° out of phase with the apply strain divided by the strain. It is a measure of the energy dissipated or lost as heat per cycle of sinusoidal deformation. Hence, the maxima on the curve  $G''(\omega)$  versus  $\omega$  are related in some way to the various molecular relaxation and thermodynamic phase transitions.

As mentioned above, for the PMR type resin, the final reaction with nadic end-groups and forms highly cross linking network required for high temperature applications. During this isothermal post curing process, the cross linking density increases. As cross linking density increases, the value of frequency-dependent  $G'(\omega)$  at a particular frequency increases. Since the post curing process is likely to be a diffusion-control process, the kinetics of network cross linking is temperature dependent. As the post curing temperature increases, the rate of network cross linking is expected to increase. Further, the rate of cross linking can be characterized with the shifting of  $G'(\omega)$  versus  $\omega$  curves obtained at

various curing times.

## **EXPERIMENTAL PROCEDURES**

### **Materials**

The preparation of the AFR700B resin samples used in this study was done by Mr. Brian Rice of University of Dayton Research Institute working at the Wright-Patterson Materials Laboratory (WP/ML). The components of the AFR700B are dimethyl 4,4'-hexfluoroisopropylidene (6-FDE), p-Phenylenediamine (p-PDA), and methyl 5-norbornene 2,3-dicarboxylate (NE). The ratio of monomers used in the AFR700B formulation is 1 : 9 : 8 of NE : p-PDA : 6-FDE respectively. The formulated average oligomeric molecular weight is thus 4380 gm/mole. The monomer mixture was imidized under vacuum at 225°C for 12 hours. The first curing reactions are induced by heating the imidized oligomer mixture at 300°C under vacuum at a hydraulic press for at least 12 hours. These samples are refer in the following as "As Cured" condition. From other study conducted at WP/ML, the "As Cured" sample exhibit a thermoplastic-like characteristics. To improve the high temperature performance of "As Cured" AFR700B resins, in practice, a post cure cycle is added. In the post cure cycle, samples are place free standing in N<sub>2</sub> atmosphere at temperature of 400 °C for times ranged up to 72 hours.

### **Environmental Degradation Condition**

In this study, "Post Cured" samples were first dried in a 100 °C vacuum oven for 24 hours to establish identical initial condition for every samples used. Samples were undergoing the hydrolytic treatment in a sealed stainless tube filled with 30 ml of water and then placed into an oven set at temperature of 200 °C for time ranged for 1 hour to 8 hours. This treatment simulating the effect of moisture on **fiber** reinforced AFR700B composites. After the hydrolytic treatment, samples were then dried in a 100 °C vacuum oven for another 24 hours to remove "free" water molecules trapped inside of the sample. The weight of each sample used was measured at the end of every stage of environmental



treatment.

### Dynamic Mechanical Spectrometry

Dynamic mechanical analysis was performed on both "AS CURED" and "ENVIRONMENTALLY DEGRADED" AFR700B resin using the Rheometry RDS-II dynamic spectrometer in the torsion rectangular fixture. To establish the initial properties, the  $G'(\omega)$  and  $G''(\omega)$  of these samples were first measured as a function of temperature from room temperature to 400 °C with heating rate of 10 °C/min. using 100 rad/second frequency and 0.2% strain. Each sample was about 47 mm x 12.5 mm with thickness of 1.3 mm. After reaching 400 °C, samples were held isothermally for a period of 16 hours. During this isothermal cure cycle, the  $G'(\omega)$  and  $G''(\omega)$  were measured as a function of frequency from 0.1 rad/sec to 100 rad/sec with strain amplitude of 0.2% at various curing times. Since there were significant increase in the cross linking density, the rate of network formation can be characterized by shifting curves of either  $G'(\omega)$  versus  $\omega$  or  $G''(\omega)$  versus  $\omega$  obtained at different cure times along the frequency axis.

## RESULTS AND DISCUSSIONS

### Initial temperature sweep

In Figure 1 we depict results of  $G'(\omega)$ ,  $G''(\omega)$  versus temperature for the initial "As Cured" and dry "hydrolytic degraded" post-cured AFR700B resin. In Figure 2 we display results of  $G'(\omega)$ ,  $G''(\omega)$  versus temperature for the "Post Cured" sample prior to any hydrolytic conditions. In compare the transition peak of  $G''$  curves, it is readily seen that upon degradation, there is significant degradation in its the glass transition. In compare the  $G'$  curves for As Cured and Post Cured, the  $G'$  curve for the As Cured sample does not exhibit a rubbery plateau regime while such regime is observed for the Post Cured sample. Further the magnitude of change when the material is to undergo transition from glassy behavior to rubbery behavior is much larger for the As Cured sample. The lack of rubbery

plateau indicates that there are little or no cross linking take place for the As Cured sample and the molecular weight of this polyimide is relatively small. Since we know the number of repeat unit for AFR700B prior to cross linking is 8. Although the peak of glass transition in the degraded sample is similar to the As Cured sample, but the width of transition is much broader. This suggest that significant scission of the post cured network take place during the hydrolysis.

### Isothermal Post Curing

In Figures 3 and 4 we display results of  $G'(\omega)$  and  $G''(\omega)$  versus frequency at different curing times for post curing a As Cured sample at 400 °C. It is readily seen the isochornal (same frequency) dynamic elastic modulus  $G'(\omega)$  increases as the curing time increase, and the dynamic modulus curve shifts to the left along the frequency axis. Applying time(frequency)-curing time superposition to such data allows one to determine the shift factor,  $a_t$ , for this data. Same shift factor can also be obtain using curves of  $G''(\omega)$  versus frequency obtained as different curing times, as shown in Figure 4. In Figure 5 we depict results of  $G'(\omega)$  versus frequency at different curing times for post curing a Hydrolytic Degraded sample at 400 °C. Similarly, the isochornal (same frequency) dynamic elastic modulus  $G'(\omega)$  increases as the curing time increase, and the dynamic modulus curve shifts to the left along the frequency axis. It is also possible to determine the shift factor for this set of data. In Figures 6, 7, and 8, the time(frequency)-curing time master curves are constructed for Figure 3, 4, and 5, respectively. To form the master curve, we use the 4 hours of post curing time as the reference time.

An important finding in comparing post curing kinetic of As Cured and Hydrolytic Degraded samples is that the shift factors needed to form master curves for the dynamic elastic and loss modulus are identical between these two samples studied. This is not a completely surprise. Since from Figure 1, we know the Hydrolytic Degraded sample has

similar glass transition as the Cured sample. Further, the network formation during the post curing cycle has been suggested to be that of diffusion control. This means the rate of network formation is control only by the temperature difference between post curing temperature and the initial glass transition of the sample.

At last in Figure 10, we show the results of  $G'(\omega)$ ,  $G''(\omega)$  versus temperature of post cured Hydrolytic Degraded sample. As indicated the glass transition peak of  $G''(\omega)$  curve shifted from 330 °C to final value of about 440 °C. The new glass transition value is similar to the post cured As Cured sample as shown in Figure 2. This suggest the stability of thermal properties, however other mechanical properties such as strength and fracture need to be examine further.

## **CONCLUSIONS**

We have shown in this study that the viscoelastic technique can be useful in the characterization of network formation during the post curing cycle of AFR700B resin. It is clear the AFR700B resin suffers significant scission of its network. However, the thermal properties can be recover by applying the high temperature post curing cycle. Other mechanical properties such as fracture behavior and ultimate strength should also be examine using hydrolytic degraded AFR700B resin and its composites after the high temperature post curing cycle. These results should be compare to a study using solid state  $^{15}\text{N}$  NM spectroscopy been conducted currently by Dr. David Curliss of WP/ML.

## **ACKNOWLEDGMENT**

I like to express my thanks to my host, Dr. David Curliss, and Mr. Brian Price of UDRI for many useful discussions. To AFOSR and RDL Summer Faculty Research Program.

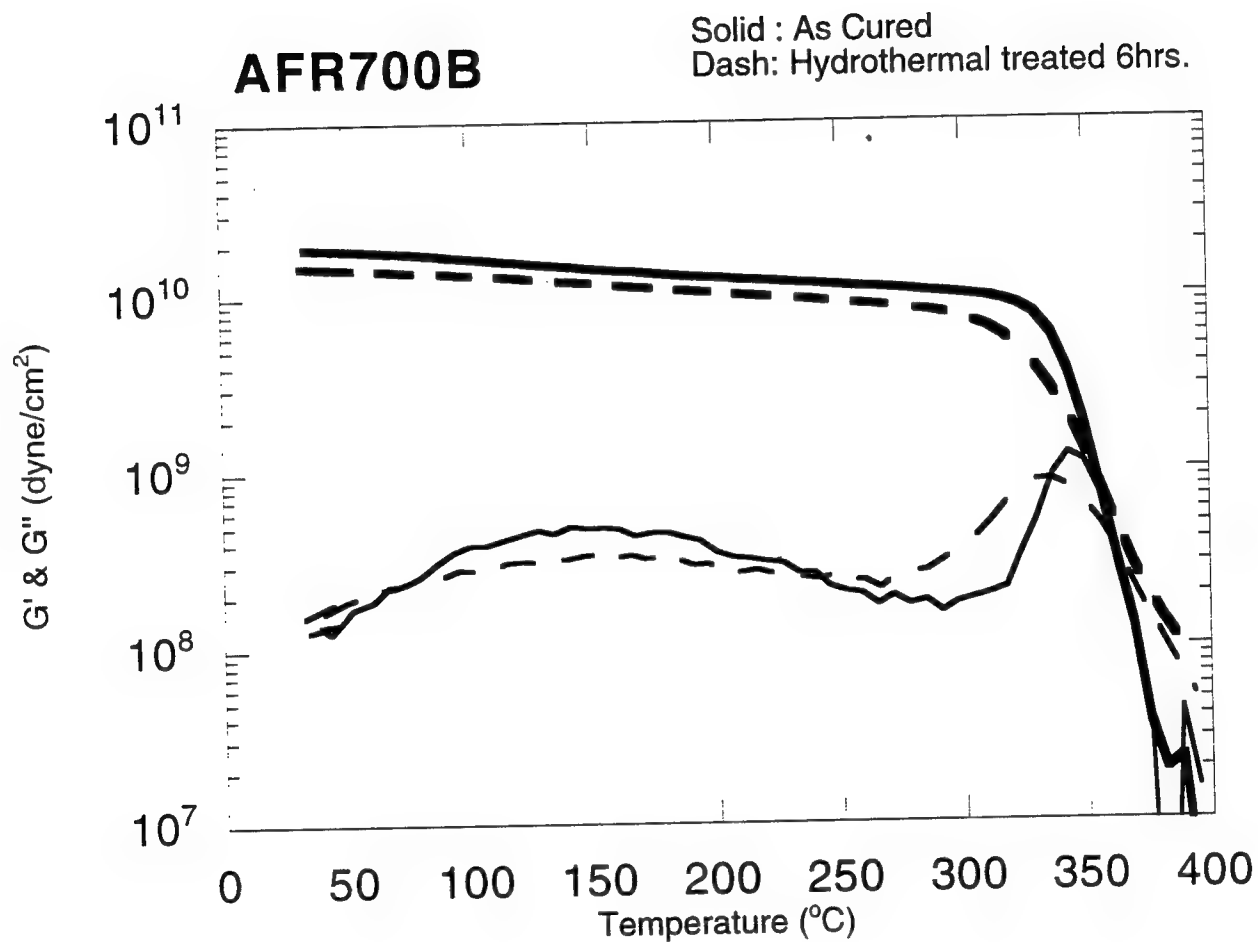


Figure 1.  $G'(\omega)$  and  $G''(\omega)$  versus temperature for As Cured and Hydrolytic Degraded Post cured AFR700B resin. The apply strain amplitude is 0.2% and frequency of 100 rad/sec. The heating rate is 10  $^{\circ}\text{C}/\text{min}$ .

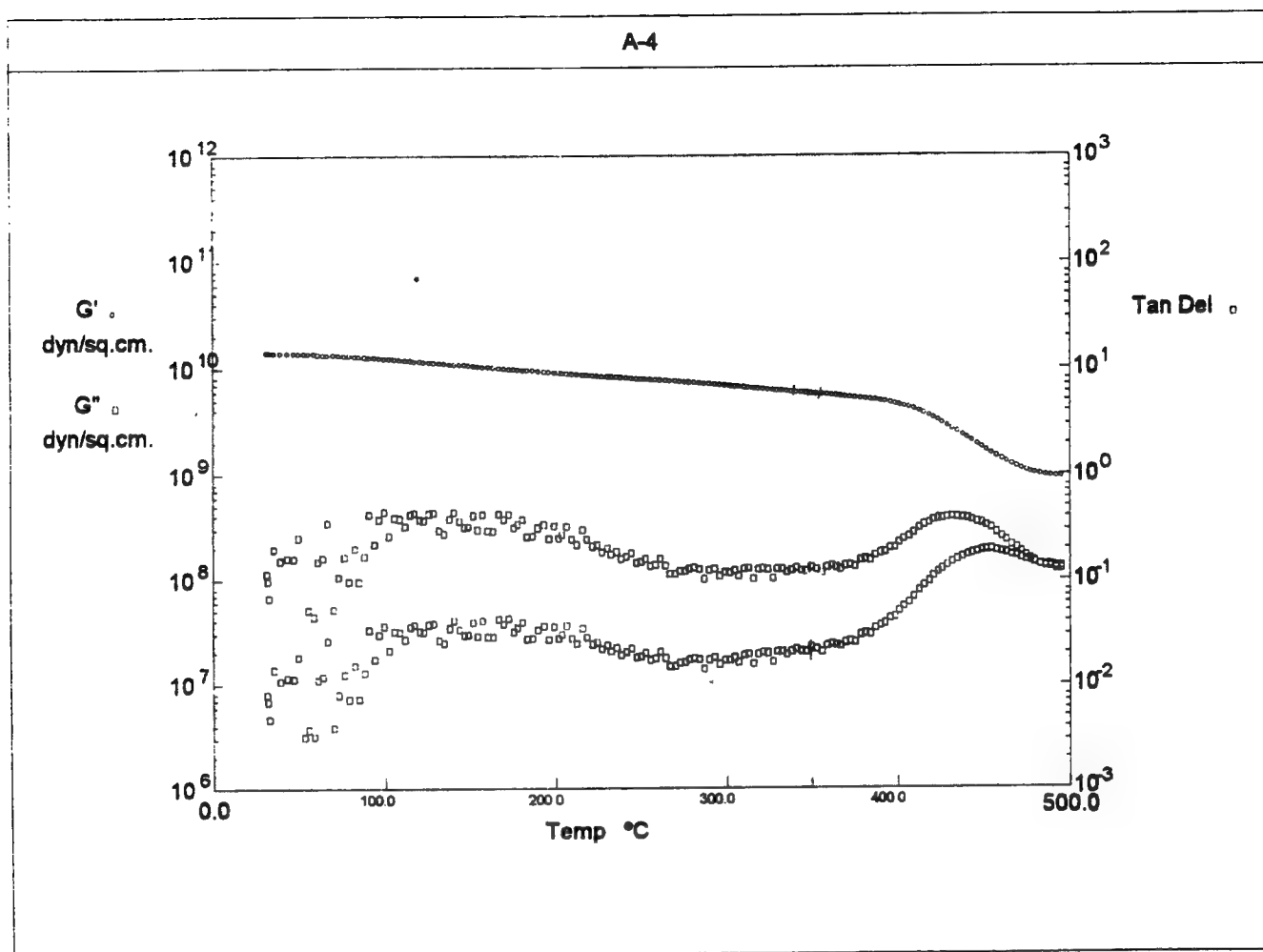


Figure 2.  $G'(\omega)$ ,  $G''(\omega)$  and  $\tan \delta$  versus temperature of a post cured AFR700B resin sample prior to the hydrolytic degradation treatment. The post curing condition is 400 °C in  $N_2$  for 16 hours. The apply strain amplitude is 0.2% and frequency of 100 rad/sec. The heating rate is 10 °C/min.

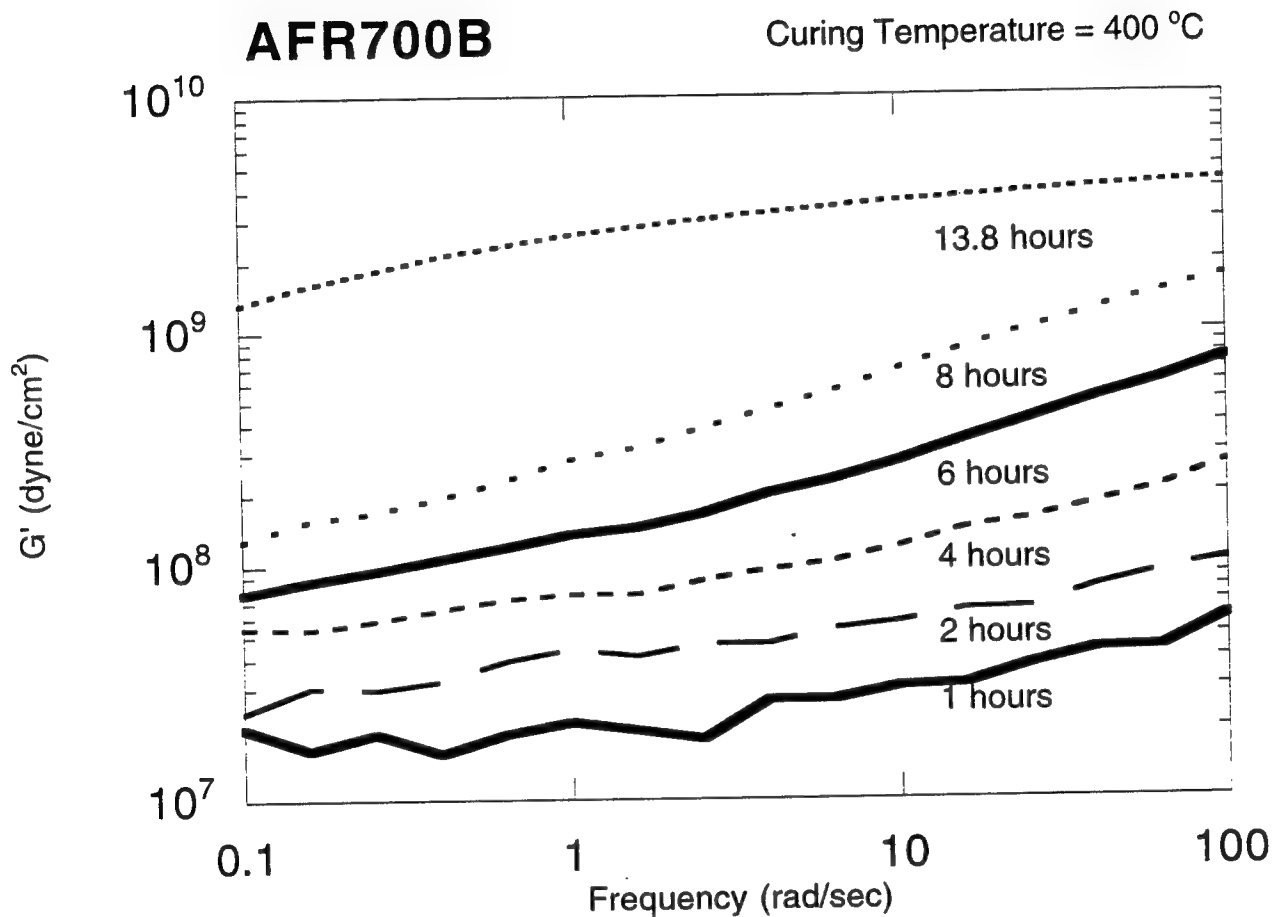


Figure 3. Post Curing of an As Cured AFR700B resin sample in the rheometer at temperature of 400 °C. The dynamic elastic modulus  $G'(\omega)$  versus frequency at different curing time as indicated is shown. The apply strain amplitude is 0.2%.

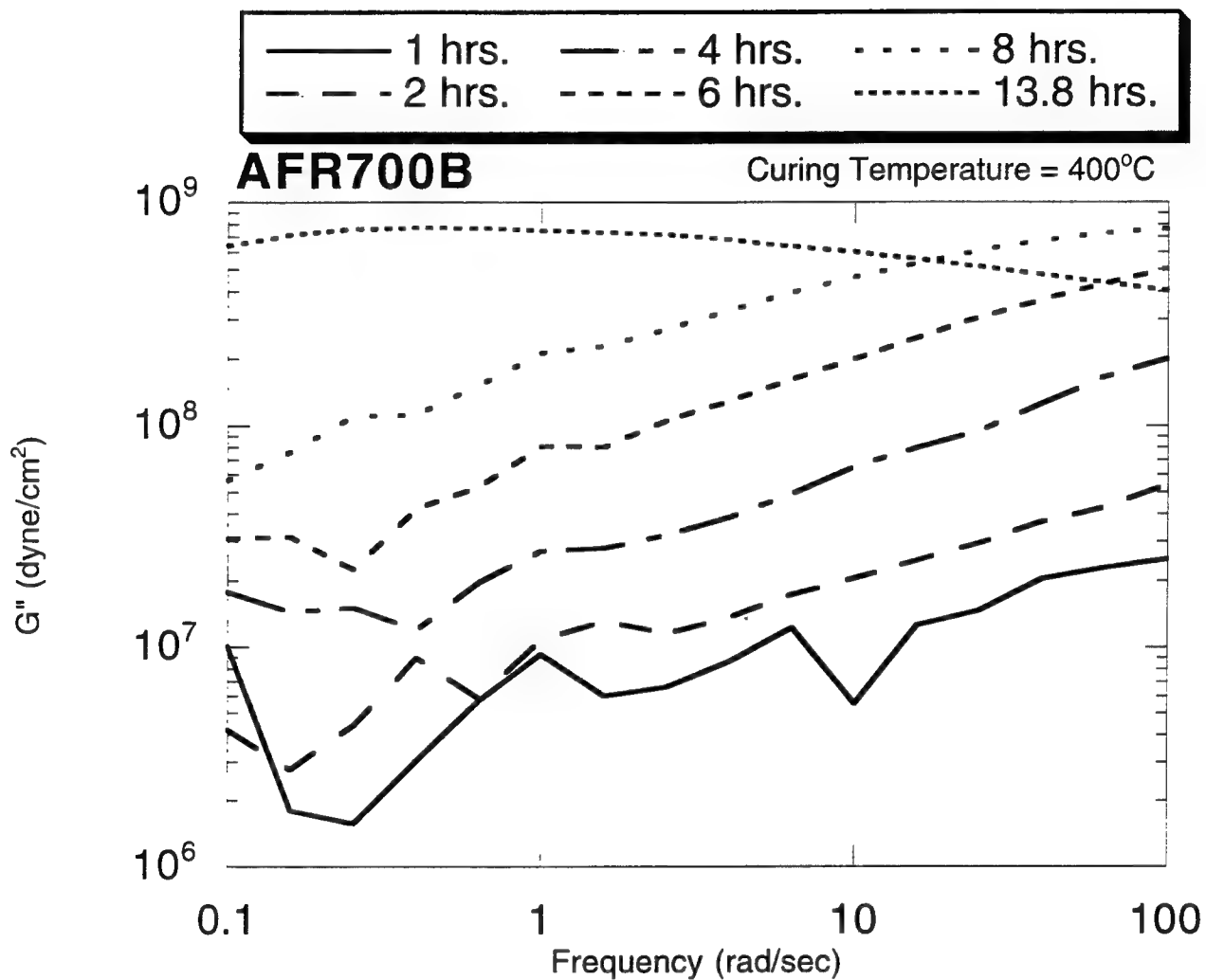


Figure 4. Post Curing of an As Cured AFR700B resin sample in the rheometer at temperature of 400 °C. The dynamic loss modulus  $G''(\omega)$  versus frequency at different curing time as indicated is shown. The apply strain amplitude is 0.2%.

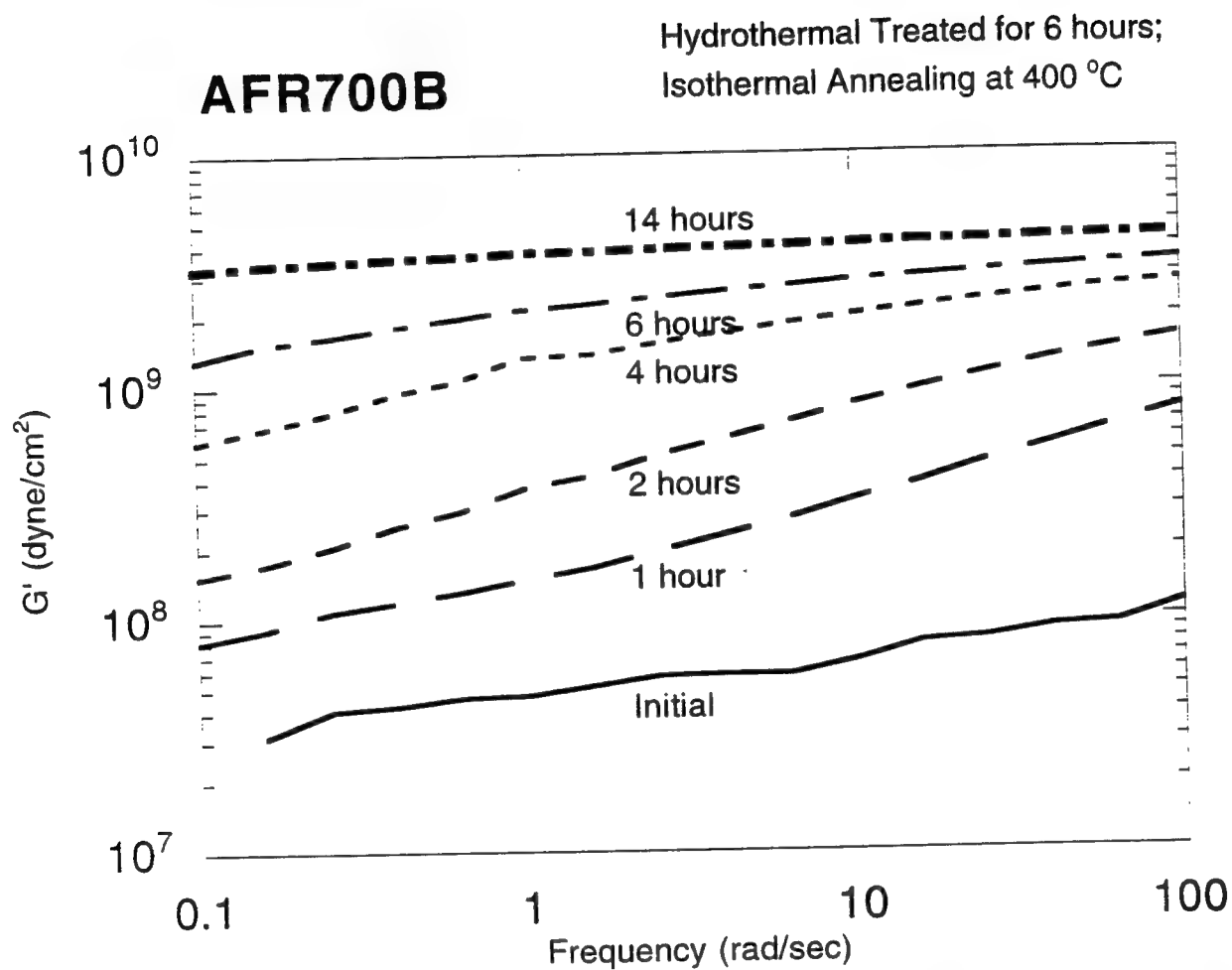


Figure 5. Post Curing of a Hydrolytic Degraded AFR700B resin sample in the rheometer at temperature of 400 °C. The dynamic elastic modulus  $G'(\omega)$  versus frequency at different curing time as indicated is shown. The apply strain amplitude is 0.2%.



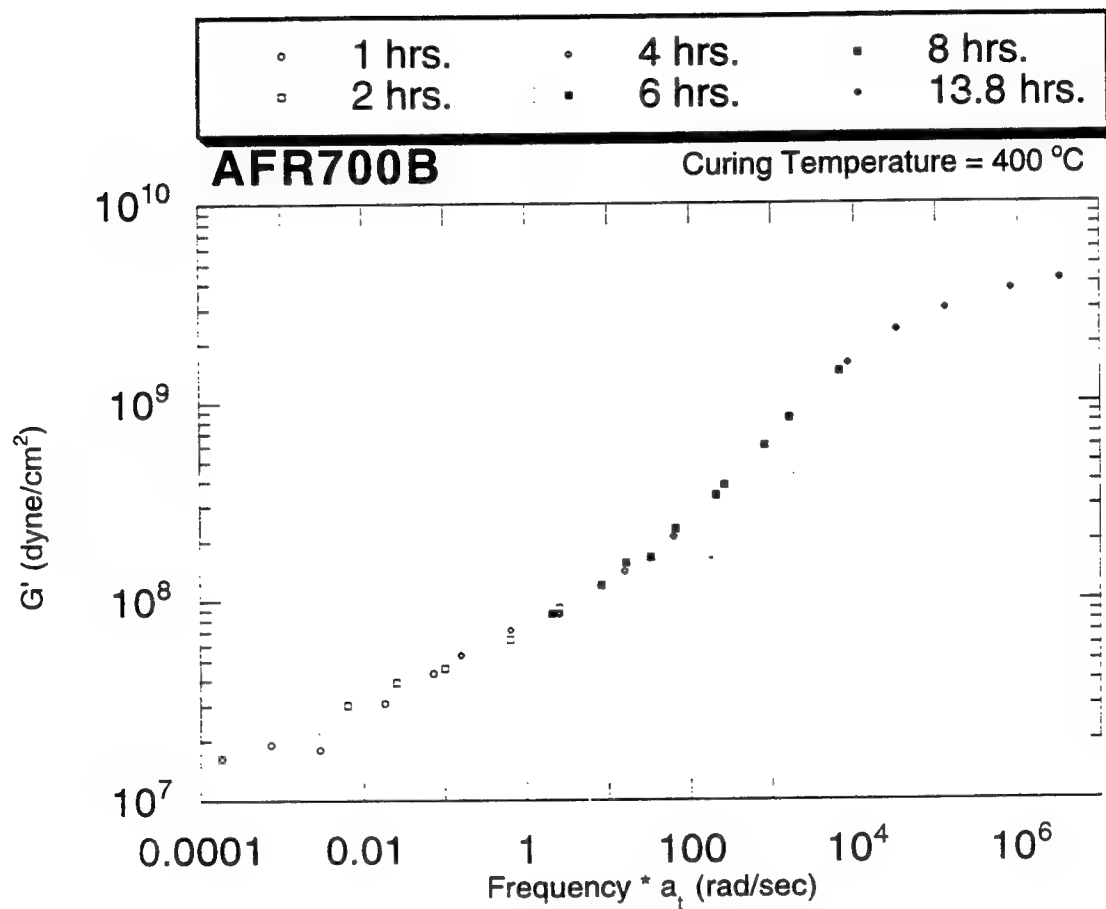


Figure 6. A master curve of time(frequency) - curing time superposition using curves shown in Figure 3. The frequency is adjust with the shift factor  $a_t$ . The reference curing time is 4 hours at 400 °C.

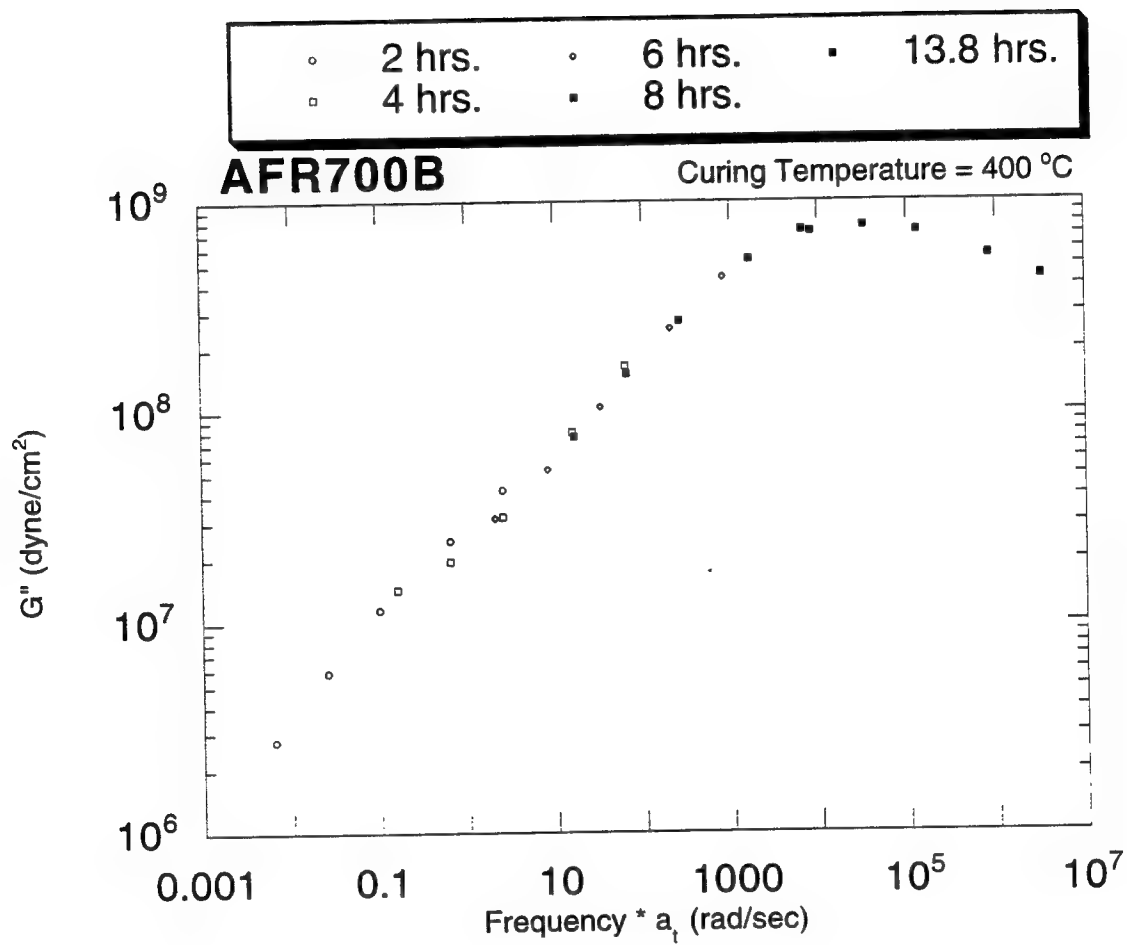


Figure 7. A master curve of time(frequency) - curing time superposition using curves shown in Figure 4. The frequency is adjust with the shift factor  $a_t$ . The reference curing time is 4 hours at 400 °C.

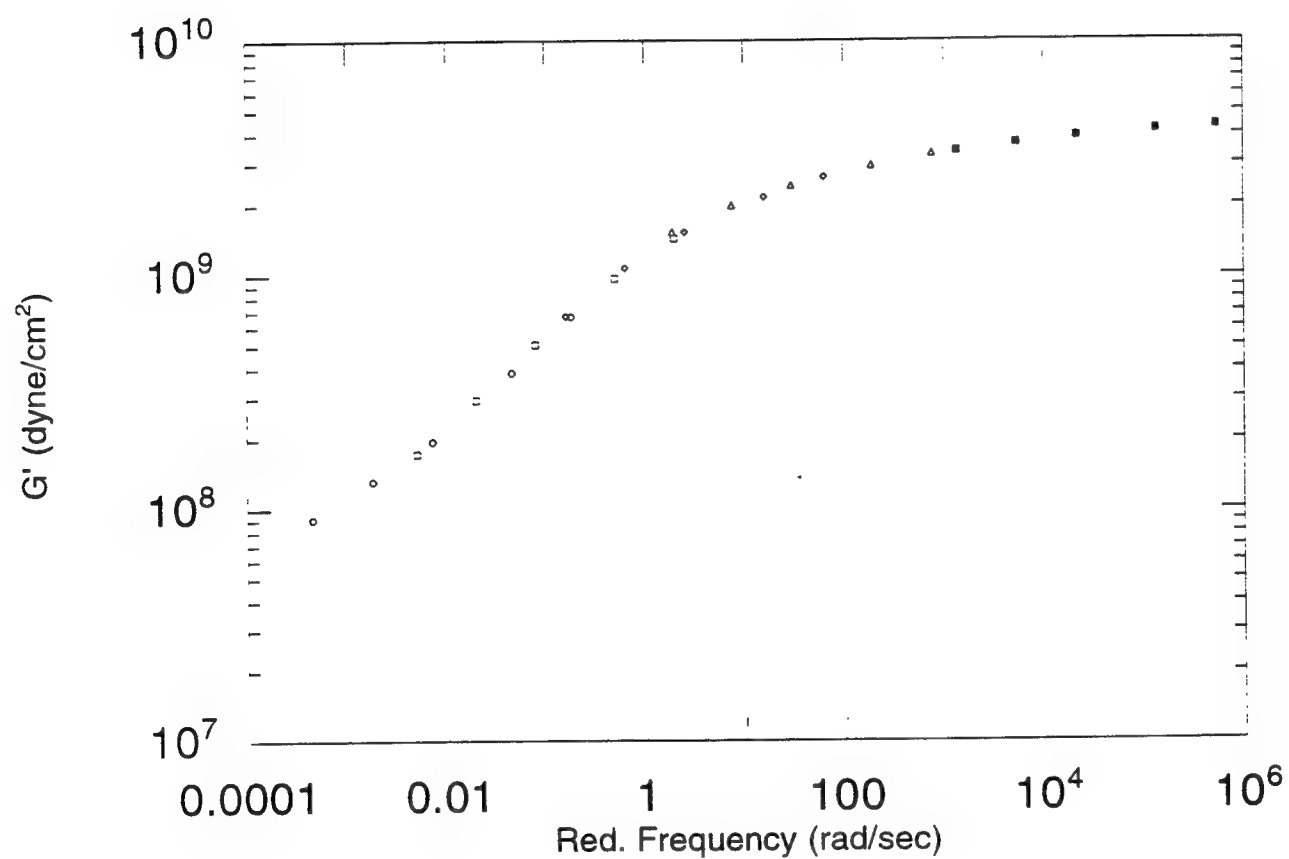


Figure 8. A master curve of time(frequency) - curing time superposition using curves shown in Figure 5. The frequency is adjust with the shift factor  $a_t$ . The reference curing time is 4 hours at 400 °C.

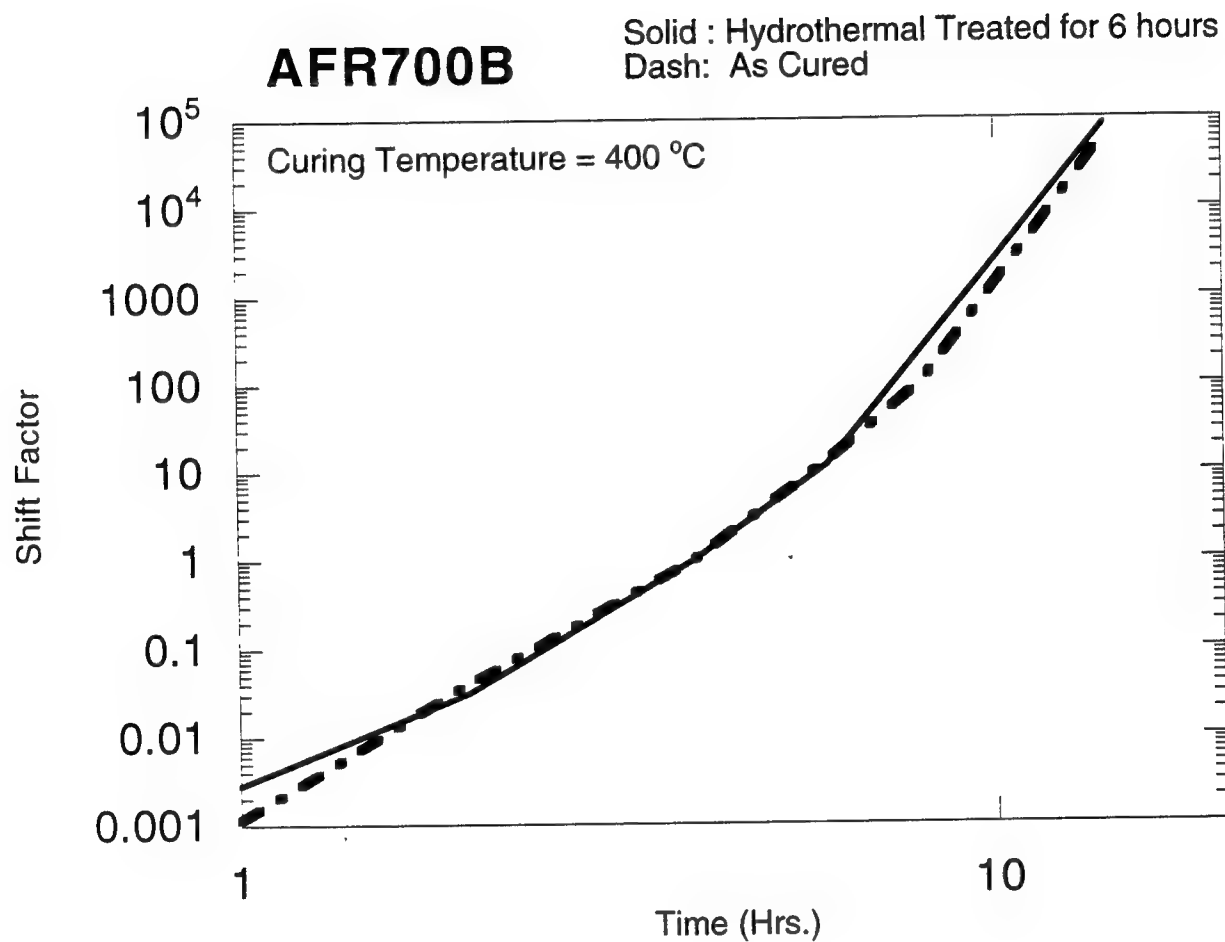


Figure 9. The time(frequency) - curing time shift factor ,  $a_t$ , used to form the master curves as shown in Figures 6 and 8 is plotted versus curing time in a log-log plot. As indicated, both As cured and Hydrolytic Degraded samples show the identical shift factor. The reference curing time is 4 hours at 400 °C.

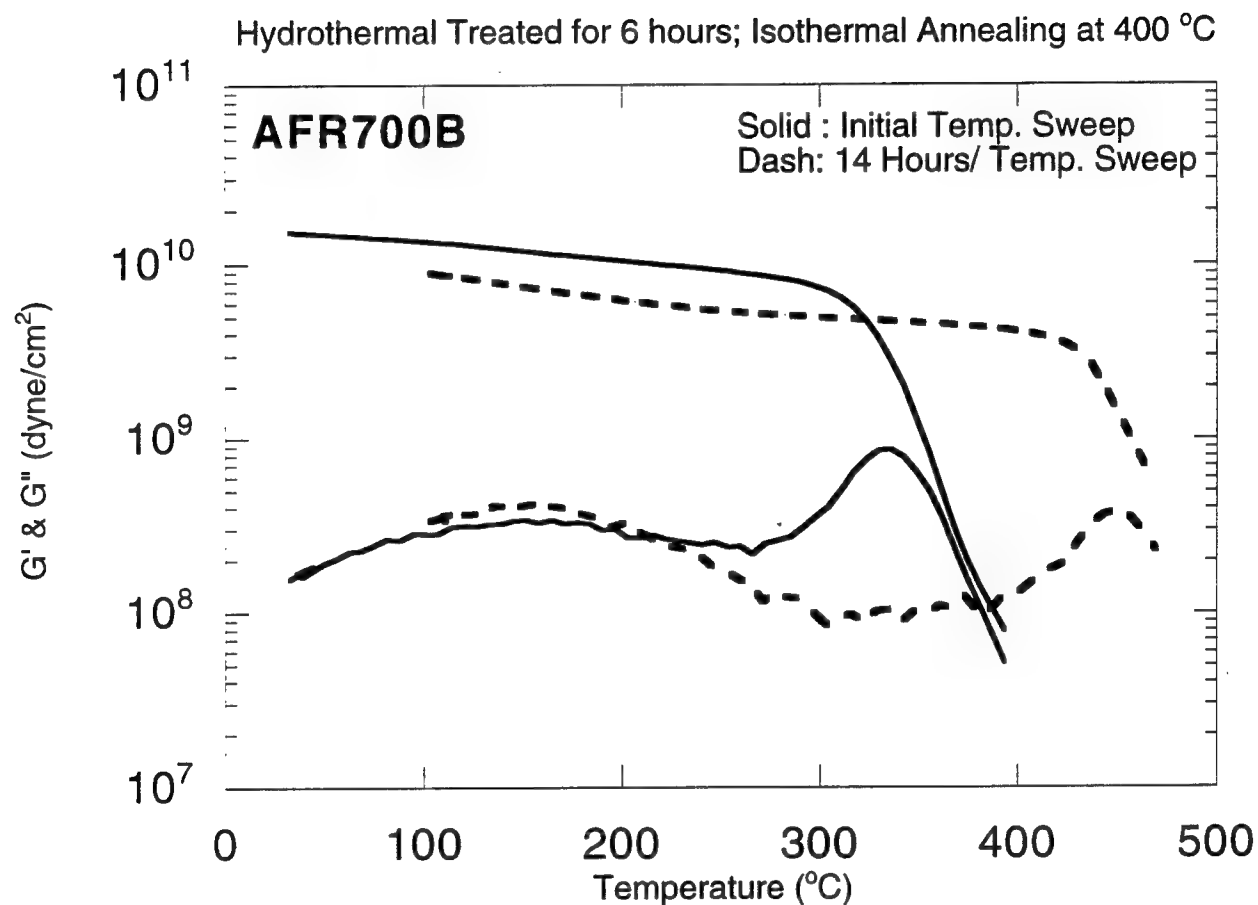


Figure 10.  $G'(\omega)$  and  $G''(\omega)$  versus temperature for the Hydrolytic Degraded sample prior to the post cure treatment and after 14 hours of post curing at 400 °C. As shown the glass transition peak in the  $G''(\omega)$  curve increased from about 330 °C to about 450 °C. The apply strain amplitude is 0.2% and frequency of 100 rad/sec. The heating rate is 10 °C/min.

OPTIMAL CONTROL AND KALMAN FILTER DESIGN FOR  
SECOND-ORDER DYNAMIC SYSTEMS

Junghsen Lieh  
Assistant Professor  
Mechanical & Materials Engineering

Wright State University  
3640 Colonel Glenn Highway  
Dayton, Ohio 45435

Final Report for:  
Summer Faculty Research Program  
Wright Laboratory

Sponsored by:  
Air Force Office of Scientific Research  
Bolling Air Force Base, DC

and

Wright Laboratory

September 1995

# OPTIMAL CONTROL AND KALMAN FILTER DESIGN FOR SECOND-ORDER DYNAMIC SYSTEMS

Junghsen Lieh  
Assistant Professor  
Mechanical & Materials Engineering  
Wright State University

## Abstract

This report describes the process of how the controllers and Kalman filters for second-order dynamic systems can be designed. Both deterministic and stochastic inputs are considered in the control system. It starts with a deterministic case where the linear quadratic regulator is developed. Both full-state and velocity-feedback control laws are discussed. The stochastic case with zero-mean excitation and measurement, which are treated as white noises, is then introduced. The optimal controller and Kalman filter are derived. The use of second-order equation for control design separates position and velocity feedback gains thus will help us understand the control mechanism. The method is further used to determine optimal velocity gains leading to savings in computing Riccati matrices. With only velocities as the feedback signal and measurement data, the data processing time can be shortened and the hardware may be simplified. An example is included, and both full-state and velocity controllers are compared with the passive system.

# OPTIMAL CONTROL AND KALMAN FILTER DESIGN FOR SECOND-ORDER DYNAMIC SYSTEMS

Junghsen Lieh

## Introduction

Structure systems can be described by second-order equations after discretization. Conventionally, control design for these systems has to convert the second-order equation into a state-space first-order form. The gain contains “mixed” states including positions and velocities as the feedback signal, which are used to generate position and velocity forces. Displacement control may require relatively high power from actuators, which potentially leads to noise and bandwidth problems. In terms of hardware and reliability, the application of position feedback to vibration suppression could be expensive. Damping control may implement sensors and control devices to change the rate of energy dissipation, or to vary the damping coefficient. The damping force can be generated from feedback velocities alone. In some applications (for example structures and vehicle suspensions), active dampers can be used to simplify the design of a control system.

The development of fully active linear controllers for an  $n$ -DOF second-order system (deterministic or stochastic) requires one to solve  $(2n^2 + n)$  algebraic Riccati equations from a converted first-order equation. In fact, the first-order Riccati equation can be expanded into three subsets such that the matrices of a second-order system may be utilized. The expansion will allow the position and velocity control forces be separated. A direct use of second-order matrices would reduce the number of “nonlinear” equations to  $(1.5n^2 + 0.5n)$ . For the case where an optimal damping controller is desired, the number of nonlinear equations can be further reduced to  $(0.5n^2 + 0.5n)$ .

The second-order method would simplify the process for controller and Kalman filter design. It allows the properties of second-order equations be preserved. A full-state system requires sensors to measure both positions and velocities and then feeds these signals back to the



actuator to generate the desired force. A damping controller only measures velocities and feeds these signals back to actuators determine the desired force. With less number of parameters to be measured and computed, damping control systems will certainly improve the computation and data processing time. With less measurement also implies that the hardware installations may be simplified.

The objective of the research work is aimed at developing optimal control algorithms that are applicable to second-order deterministic and stochastic second-order systems. The paper starts with a deterministic case for linear quadratic regulator problems. It then extends to a stochastic system with white noise excitations and measurements. Both full-state and pure velocity feedback systems are considered. The Kalman filter for second-order systems is then introduced. The use of second-order equations for control design decomposes position and velocity gains, which will help us understand the control mechanism. It can be used to derive optimal damping controllers potentially leading to savings in computing and data processing time. A simple example is used in this report to illustrate the procedure of how a controller is determined.

## **Methodology**

This section describes the procedure of control design that will lead to optimal gains and Kalman filters. By using second-order equations, the method separates the feedback forces generated by positions and velocities. With only velocities as the feedback signal, the effort for computing gain matrices and Kalman filters is reduced. The reduced controller requires less data processing time in the feedback loop. Two sub-sections are presented. The first is for deterministic systems and the second is for systems with stochastic inputs.

### **1. Controller for Deterministic Systems**

Consider a second-order dynamic system

$$M\ddot{y} + D\dot{y} + Ky = bu \quad (1.1)$$

$$\underline{z} = \underline{C}_1 \underline{y} + \underline{C}_2 \dot{\underline{y}} \quad (1.2)$$

The control system is intended to minimize a performance index

$$J = \int_{t_0}^{t_f} (\underline{y}^T \underline{Q}_{11} \underline{y} + 2 \underline{y}^T \underline{Q}_{12} \dot{\underline{y}} + \dot{\underline{y}}^T \underline{Q}_{22} \dot{\underline{y}} + 2 \underline{y}^T \underline{S}_1^T \underline{u} + 2 \dot{\underline{y}}^T \underline{S}_2^T \underline{u} + \underline{u}^T \underline{R} \underline{u}) dt \quad (1.3)$$

Define

$$\underline{u} = \hat{\underline{u}} - \underline{R}^{-1}(\underline{S}_1 \underline{y} + \underline{S}_2 \dot{\underline{y}}) \quad (1.4)$$

in which the control weighting matrix  $\underline{R}$  is assumed to be nonsingular. Substituting Eqn (1.4) into Eqn (1.1) leads to

$$\underline{M} \ddot{\underline{y}} + \underline{\mathcal{D}} \dot{\underline{y}} + \underline{\mathcal{K}} \underline{y} = \underline{b} \hat{\underline{u}} \quad (1.5)$$

which is equivalent to minimizing

$$J = \int_{t_0}^{t_f} (\underline{y}^T \underline{\mathcal{Q}}_{11} \underline{y} + 2 \underline{y}^T \underline{\mathcal{Q}}_{12} \dot{\underline{y}} + \dot{\underline{y}}^T \underline{\mathcal{Q}}_{22} \dot{\underline{y}} + \hat{\underline{u}}^T \underline{R} \hat{\underline{u}}) dt \quad (1.6)$$

Define  $\underline{P}_{11}$ ,  $\underline{P}_{12}$ , and  $\underline{P}_{22}$  as the Riccati submatrices

$$\underline{P} = \begin{bmatrix} \underline{P}_{11} & \underline{P}_{12} \\ \underline{P}_{12}^T & \underline{P}_{22} \end{bmatrix} \quad (1.7)$$

of the converted first-order equation, i.e.,

$$\dot{\underline{x}} = \underline{A} \underline{x} + \underline{B} \hat{\underline{u}} \quad (1.8)$$

where  $\underline{x} = [\underline{y}^T, \dot{\underline{y}}^T]^T$  and

$$\underline{A} = \begin{bmatrix} \underline{0} & \underline{I} \\ -\underline{M}^{-1} \underline{\mathcal{K}} & -\underline{M}^{-1} \underline{\mathcal{D}} \end{bmatrix} \quad (1.9)$$

$$\underline{B} = \begin{bmatrix} \underline{0} \\ \underline{M}^{-1} \underline{b} \end{bmatrix} \quad (1.10)$$

Through a matrix expansion, the controller for Eqn (1.5) becomes

$$\underline{\dot{u}} = -R^{-1}(\mathcal{M}^1 \underline{b})^T \mathcal{P}_{12}^T \underline{x} - R^{-1}(\mathcal{M}^1 \underline{b})^T \mathcal{P}_{22} \underline{\dot{y}} \quad (1.11)$$

and the Riccati equations for the second-order system are

$$\mathcal{P}_{12}(\mathcal{M}^1 \underline{b})R^{-1}(\mathcal{M}^1 \underline{b})^T \mathcal{P}_{12}^T + \mathcal{P}_{12}(\mathcal{M}^1 \underline{\mathcal{K}}) + (\mathcal{M}^1 \underline{\mathcal{K}})^T \mathcal{P}_{12}^T - \underline{\mathcal{Q}}_{11} = 0 \quad (1.12a)$$

$$\mathcal{P}_{22} \mathcal{M}^1 \underline{b} R^{-1}(\mathcal{M}^1 \underline{b})^T \mathcal{P}_{22} + \mathcal{P}_{22}(\mathcal{M}^1 \underline{\mathcal{D}}) + (\mathcal{M}^1 \underline{\mathcal{D}})^T \mathcal{P}_{22} - \mathcal{P}_{12} - \mathcal{P}_{12}^T - \underline{\mathcal{Q}}_{22} = 0 \quad (1.12b)$$

$$\mathcal{P}_{11} = \mathcal{P}_{12}(\mathcal{M}^1 \underline{\mathcal{D}}) + (\mathcal{M}^1 \underline{\mathcal{K}})^T \mathcal{P}_{22} + \mathcal{P}_{12}(\mathcal{M}^1 \underline{b})R^{-1}(\mathcal{M}^1 \underline{b})^T \mathcal{P}_{22} - \underline{\mathcal{Q}}_{12} \quad (1.12c)$$

The control law for a full-state feedback system is expressed as follows:

$$\underline{u} = -R^{-1}[(\mathcal{M}^1 \underline{b})^T \mathcal{P}_{12}^T + \mathcal{S}_1] \underline{x} - R^{-1}[(\mathcal{M}^1 \underline{b})^T \mathcal{P}_{22} + \mathcal{S}_2] \underline{\dot{y}} \quad (1.13)$$

It can be seen that for a full-state feedback system, only  $\mathcal{P}_{12}$  and  $\mathcal{P}_{22}$  are needed for the controller. Due to coupling effects, it will be necessary to solve these Riccati sub-equations simultaneously in order to determine  $\mathcal{P}_{11}$ ,  $\mathcal{P}_{12}$  and  $\mathcal{P}_{22}$  if the first-order method is used. Under this case, the number of unknowns in  $\mathcal{P}_{ij}$  will be  $(2n^2 + n)$ . However, if the second-order method is used, the submatrix  $\mathcal{P}_{12}$  can be determined using Eqn (1.12a) alone. It then determines  $\mathcal{P}_{22}$  by substituting  $\mathcal{P}_{12}$  into Eqn (1.12b). It is obvious that the number of nonlinear unknowns is reduced to  $(1.5n^2 + 0.5n)$ , leading to a saving of nearly 25%. It should be noted that to compute a non-symmetric  $\mathcal{P}_{12}$ , a general procedure will be needed.

The control law described above is developed with both position and velocity feedback signals. For certain control systems, it may not be necessary to include all feedback information. Typical example is vibration suppression in which the use of position feedback could lead to complexity in design and operation, particularly for a large-scale system. In such a case, the use of velocity feedback may be considered. With velocities as the only feedback signal, the hardware as well as software could be simplified, and the data processing time may be improved.

Simplifying Eqn (1.13), the control law for a velocity feedback system can be written as

$$\underline{u} = -R^{-1}[(M^1 b)^T P_{22} + S_2] \dot{\underline{y}} \quad (1.14)$$

The velocity control system is equivalent to imposing  $(M^1 b)^T P_{12}^T + S_1 = 0$ . To allow the velocity control be acceptable, the following condition must be satisfied: If the position is not used as a performance measure (i.e.,  $Q_{11} = 0$  and  $S_1 = 0$ ), one solution to the first partitioned equation is  $P_{12} = 0$ . It is observed that the submatrix  $P_{22}$  can be directly obtained from Eqn (1.12b), i.e.,

$$P_{22}(M^1 b)R^{-1}(M^1 b)^T P_{22} + P_{22}(M^1 \underline{d}) + (M^1 \underline{d})^T P_{22} - \underline{z}_{22} = 0 \quad (1.15)$$

Since  $P_{22}$  is a Riccati matrix, existing procedures available in MATLAB and Matrix<sub>x</sub> can be used. The simplification reduces the number of unknowns from previous  $(2n^2 + n)$  to  $(0.5n^2 + 0.5n)$ , which is nearly a 4:1 ratio for large  $n$ . To have the velocity feedback system controllable, the controllability test matrix must be fully ranked.

## 2. Control Design for Stochastic Inputs

Consider a second-order system

$$M \ddot{\underline{y}} + D \dot{\underline{y}} + K \underline{y} = b \underline{u} + f \underline{v} \quad (2.1)$$

$$\underline{z} = \underline{C}_1 \underline{y} + \underline{C}_2 \dot{\underline{y}} + \underline{w} \quad (2.2)$$

where  $\underline{v}$  and  $\underline{w}$  are zero-mean white noises with spectral density  $\underline{V}$  and  $\underline{W}$ , i.e.,

$$E\{\underline{v}\} = 0; \quad E\{\underline{v}(t) \underline{v}(\tau)^T\} = \underline{V} \delta(t-\tau) \quad (2.3a)$$

$$E\{\underline{w}\} = 0; \quad E\{\underline{w}(t) \underline{w}(\tau)^T\} = \underline{W} \delta(t-\tau) \quad (2.3b)$$

where the symbol  $E\{\}$  denotes the mathematical expectation. The control system is intended to minimize a performance index

$$J = E \left\{ \int_{t_0}^t (\underline{y}^T Q_{11} \underline{y} + 2 \underline{y}^T Q_{12} \dot{\underline{y}} + \dot{\underline{y}}^T Q_{22} \dot{\underline{y}} + 2 \underline{y}^T S_1^T \underline{u} + 2 \dot{\underline{y}}^T S_2^T \underline{u} + \underline{u}^T R \underline{u}) dt \right\} \quad (2.4)$$

Use the same procedure as described for the deterministic case, the performance index becomes

$$J = E \left\{ \int_{t_0}^t (\underline{y}^T Q_{11} \underline{y} + 2 \underline{y}^T Q_{12} \dot{\underline{y}} + \dot{\underline{y}}^T Q_{22} \dot{\underline{y}} - \underline{y}^T S_1^T R^{-1} S_1 \underline{y} - 2 \underline{y}^T S_1^T R^{-1} S_2 \dot{\underline{y}} - \dot{\underline{y}}^T S_2^T R^{-1} S_2 \dot{\underline{y}} + \dot{\underline{u}}^T R \dot{\underline{u}}) dt \right\} \quad (2.5)$$

Substituting the controller Eqn (1.4) into Eqn (2.1) leads to

$$M \ddot{\underline{y}} + \underline{D} \dot{\underline{y}} + \underline{K} \underline{y} = \underline{b} \dot{\underline{u}} + \underline{f} \underline{y} \quad (2.6)$$

which is equivalent to minimizing

$$J = E \left\{ \int_{t_0}^t (\underline{y}^T \underline{Q}_{11} \underline{y} + 2 \underline{y}^T \underline{Q}_{12} \dot{\underline{y}} + \dot{\underline{y}}^T \underline{Q}_{22} \dot{\underline{y}} + \dot{\underline{u}}^T R \dot{\underline{u}}) dt \right\} \quad (2.7)$$

$$= \underline{y}(t_0)^T P_{11}(t_0, t_f) \underline{y}(t_0) + 2 \underline{y}(t_0)^T P_{12}(t_0, t_f) \dot{\underline{y}}(t_0) + \dot{\underline{y}}(t_0)^T P_{22}(t_0, t_f) \dot{\underline{y}}(t_0) + \int_{t_0}^t \text{tr}[(M^1 \dot{\underline{f}})^T \underline{y} (M^1 \dot{\underline{f}})^T P_{22}(t, t_f)] dt \quad (2.8)$$

Expand from the corresponding first-order equation

$$\dot{\underline{x}} = \underline{A} \underline{x} + \underline{B} \dot{\underline{u}} + \underline{E} \underline{y} \quad (2.9)$$

where  $\underline{A}$  and  $\underline{B}$  have been defined previously and

$$\underline{E} = \begin{bmatrix} \underline{0} \\ \underline{M}^{-1} \underline{f} \end{bmatrix} \quad (2.10)$$

the controller for Eqn (2.6) is same as Eqns (1.11) and (1.13), i.e.,

$$\dot{\underline{u}} = -R^{-1}(\underline{M}^1 \underline{b})^T P_{12}^T \underline{y} - R^{-1}(\underline{M}^1 \underline{b})^T P_{22} \dot{\underline{y}}$$

and

$$\underline{u} = -R^{-1}[(\underline{M}^1 \underline{b})^T P_{12}^T + S_1] \underline{y} - R^{-1}[(\underline{M}^1 \underline{b})^T P_{22} + S_2] \dot{\underline{y}}$$

The Riccati submatrices  $P_{12}$  and  $P_{22}$  are obtained from the same procedure as described in Eqn (1.12).

Substituting the control  $u$  into Eqn (2.1) yields a closed-loop system

$$\begin{aligned} M \ddot{y} + \{D + b R^{-1}[(M^1 b)^T P_{22} + S_2]\} \dot{y} \\ + \{K + b R^{-1}[(M^1 b)^T P_{12}^T + S_1]\} y = f y \end{aligned} \quad (2.11)$$

The power spectrum for the closed-loop system is given as

$$S_y(\omega) = H(-j\omega) V H^T(j\omega) \quad (2.12)$$

where  $H(j\omega) = Y(j\omega)/V(j\omega)$  is the transfer function of the closed-loop system, i.e.,

$$\begin{aligned} H(j\omega) = \{K + b R^{-1}[(M^1 b)^T P_{12}^T + S_1] - \omega^2 M \\ + j\omega \{D + b R^{-1}[(M^1 b)^T P_{22} + S_2]\}\}^{-1} f \end{aligned} \quad (2.13)$$

With velocities as the only feedback signal, the controller is same as Eqn (1.14), i.e.,

$$u = -R^{-1}[(M^1 b)^T P_{22} + S_2] \dot{y}$$

and  $P_{22}$  can be directly obtained from Eqn (1.15), or from the following expanded equation

$$\begin{aligned} P_{22}(M^1 b)R^{-1}(M^1 b)^T P_{22} + P_{22}[M^1(D + b R^{-1}S_2)] \\ + [M^1(D + b R^{-1}S_2)]^T P_{22} - (Q_{22} - S_2^T R^{-1}S_2) = 0 \end{aligned} \quad (2.14)$$

Substituting the damping control force into Eqn (2.1) yields the closed-loop system in the following form

$$M \ddot{y} + \{D + b R^{-1}[(M^1 b)^T P_{22} + S_2]\} \dot{y} + K y = f y \quad (2.15)$$

The power spectrum of  $y$  is represented by

$$S_y(\omega) = H(-j\omega) V H^T(j\omega)$$

where the transfer function is defined by

$$H(j\omega) = \{K - \omega^2 M + j\omega \{D + b R^{-1}[(M^{-1}b)^T P_{22} + S_2]\}\}^{-1} f \quad (2.16)$$

### 3. Kalman Filter

In addition to the system dynamics, an observer may be estimated by

$$M \ddot{\underline{y}} + D \dot{\underline{y}} + K \underline{y} = b u + k (z - C_1 \underline{y} - C_2 \dot{\underline{y}}) \quad (3.1)$$

Define the estimation error by

$$\underline{e} = \underline{y} - \underline{\hat{y}} \quad (3.2)$$

Subtracting Eqn (3.1) from Eqn (2.1) yields an error dynamics

$$M \ddot{\underline{e}} + (D + k C_2) \dot{\underline{e}} + (K + k C_1) \underline{e} = f \underline{v} - k \underline{w} \quad (3.3)$$

Expand the Kalman filter and the associated Riccati equation, i.e.,

$$\underline{\hat{K}} = L C^T W^{-1} \quad (3.4)$$

$$A L + L A^T - L C^T W^{-1} C L + E V E^T = 0 \quad (3.5)$$

for the corresponding first-order system yields

$$(L_{11} C_1^T + L_{12} C_2^T) W^{-1} (C_1 L_{11} + C_2 L_{12}) - L_{12}^T - L_{12} = 0 \quad (3.6a)$$

$$(L_{11} C_1^T + L_{12} C_2^T) W^{-1} (C_1 L_{12} + C_2 L_{22}) - L_{22} + L_{11} (M^{-1} K)^T + L_{12} (M^{-1} D)^T = 0 \quad (3.6b)$$

$$(L_{12}^T C_1^T + L_{22} C_2^T) W^{-1} (C_1 L_{12} + C_2 L_{22}) + (M^{-1} K) L_{12} + (M^{-1} D) L_{22} \\ + L_{12}^T (M^{-1} K)^T + L_{22} (M^{-1} D)^T - (M^{-1} f) V (M^{-1} f)^T = 0 \quad (3.6c)$$

It is equivalent to having

$$\underline{\hat{k}}_1 = (L_{11} C_1^T + L_{12} C_2^T) W^{-1} \quad (3.7)$$

$$\underline{\hat{k}}_2 = (L_{12}^T C_1^T + L_{22} C_2^T) W^{-1} \quad (3.8)$$

If only velocities are needed and measured (i.e.,  $\underline{C}_1 = 0$ ) then the output becomes

$$\underline{z} = \underline{C}_2 \dot{\underline{x}} + \underline{w} \quad (3.9)$$

Equation (3.6) can be simplified as follows:

$$\underline{L}_{12} \underline{C}_2^T \underline{W}^{-1} \underline{C}_2 \underline{L}_{12}^T - \underline{L}_{12}^T - \underline{L}_{12} = \underline{0} \quad (3.10a)$$

$$\underline{L}_{12} \underline{C}_2^T \underline{W}^{-1} \underline{C}_2 \underline{L}_{22} - \underline{L}_{22} + \underline{L}_{11} (\underline{M}^{-1} \underline{K})^T + \underline{L}_{12} (\underline{M}^{-1} \underline{D})^T = \underline{0} \quad (3.10b)$$

$$\begin{aligned} \underline{L}_{22} \underline{C}_2^T \underline{W}^{-1} \underline{C}_2 \underline{L}_{22} + (\underline{M}^{-1} \underline{K}) \underline{L}_{12} + (\underline{M}^{-1} \underline{D}) \underline{L}_{22} + \underline{L}_{12}^T (\underline{M}^{-1} \underline{K})^T \\ + \underline{L}_{22} (\underline{M}^{-1} \underline{D})^T - (\underline{M}^{-1} \underline{f}) \underline{V} (\underline{M}^{-1} \underline{f})^T = \underline{0} \end{aligned} \quad (3.10c)$$

One solution to Eqn (3.10a) will be  $\underline{L}_{12} = \underline{0}$ . The Kalman filter is reduced to

$$\tilde{\underline{k}}_2 = \underline{L}_{22} \underline{C}_2^T \underline{W}^{-1} \quad (3.11)$$

with  $\underline{L}_{22}$  being computed from the simplified form of Eqn (3.10c)

$$\underline{L}_{22} \underline{C}_2^T \underline{W}^{-1} \underline{C}_2 \underline{L}_{22} + (\underline{M}^{-1} \underline{D}) \underline{L}_{22} + \underline{L}_{22} (\underline{M}^{-1} \underline{D})^T - (\underline{M}^{-1} \underline{f}) \underline{V} (\underline{M}^{-1} \underline{f})^T = \underline{0} \quad (3.12)$$

### Example

Consider a 1-DOF simple model as shown in Figure 1.

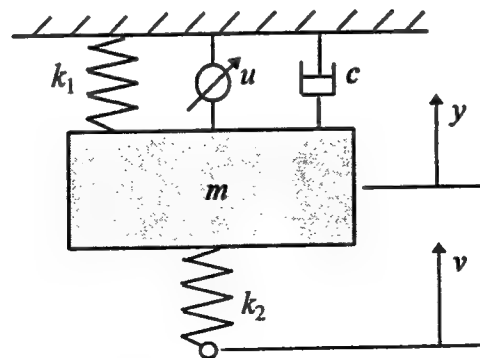


Figure 1. Schematic of a one-DOF system.



The equation of motion can be expressed by

$$m \ddot{y} + c \dot{y} + (k_1 + k_2)y = u + k_2 v \quad (4.1)$$

where

$m$  = mass

$c$  = damping coefficient

$k_i$  = spring constant,  $i = 1, 2$

$u$  = control force

$y$  = displacement

$v$  = excitation

If the first-order method is used, it will be necessary to convert the second-order equation into a first-order form, i.e.,

$$\dot{\underline{x}} = \underline{A} \underline{x} + \underline{B} u + \underline{E} v \quad (4.2)$$

where

$$\underline{A} = \begin{bmatrix} 0 & 1 \\ \frac{-(k_1 + k_2)}{m} & -\frac{c}{m} \end{bmatrix}$$

$$\underline{B} = \begin{bmatrix} 0 & \frac{1}{m} \end{bmatrix}^T$$

$$\underline{E} = \begin{bmatrix} 0 & \frac{k_2}{m} \end{bmatrix}^T$$

However, by using second-order approaches, this conversion procedure is omitted. The following sections describe how these three controllers are derived. It starts with the conventional first-order method, then followed by the full-state second-order scheme, and finally the velocity feedback system.

**Case 1: First-order full state feedback**

Performance index:

$$J = E \left\{ \int_{t_0}^{t_f} (\mathbf{x}^T \mathbf{Q} \mathbf{x} + 2 \mathbf{x}^T \mathbf{S}^T \mathbf{u} + r u^2) dt \right\}$$

Steady-state Riccati equations:

$$\mathbf{P}(\mathbf{A} - \mathbf{B} \mathbf{R}^{-1} \mathbf{S}) + (\mathbf{A} - \mathbf{B} \mathbf{R}^{-1} \mathbf{S})^T \mathbf{P} - \mathbf{P} \mathbf{B} \mathbf{R}^{-1} \mathbf{B}^T \mathbf{P} + \mathbf{Q} - \mathbf{S}^T \mathbf{R}^{-1} \mathbf{S} = \mathbf{0} \quad (4.3)$$

$$\frac{1}{m^2 r} p_{12}^2 + \frac{2}{m} \left( k_1 + k_2 + \frac{s_1}{r} \right) p_{12} - q_{11} + \frac{s_1^2}{r} = 0 \quad (4.4a)$$

$$p_{11} - \frac{1}{m^2 r} p_{12} p_{22} - \frac{1}{m} \left( c + \frac{s_2}{r} \right) p_{12} - \frac{1}{m} \left( k_1 + k_2 + \frac{s_1}{r} \right) p_{22} + q_{12} - \frac{s_1 s_2}{r} = 0 \quad (4.4b)$$

$$\frac{1}{m^2 r} p_{22}^2 + \frac{2}{m} \left( c + \frac{s_2}{r} \right) p_{22} - 2p_{12} - q_{22} + \frac{s_2^2}{r} = 0 \quad (4.4c)$$

Controller:

$$\mathbf{u} = -\mathbf{R}^{-1}(\mathbf{B}^T \mathbf{P} + \mathbf{S}) \mathbf{x} \quad (4.5)$$

When using the first-order approach, the variables  $p_{11}$ ,  $p_{12}$ , and  $p_{22}$  are computed simultaneously from the above nonlinear Riccati equations. Although  $p_{11}$  can be substituted after  $p_{12}$  and  $p_{22}$  have been solved, the first-order approach is unable to use this advantage.

**Case 2: Second-order full state feedback**

Performance index:

$$J = E \left\{ \int_{t_0}^{t_f} (q_1 \dot{y}^2 + 2 q_{12} \dot{y} \dot{y} + q_{22} \dot{y}^2 + 2 s_1 y u + 2 s_2 \dot{y} u + r u^2) dt \right\} \quad (4.6)$$

Steady-state Riccati equations:

$$\frac{1}{m^2 r} p_{12}^2 + \frac{2}{m} \left( k_1 + k_2 + \frac{s_1}{r} \right) p_{12} - q_{11} + \frac{s_1^2}{r} = 0 \quad (4.7a)$$

$$\frac{1}{m^2 r} p_{22}^2 + \frac{2}{m} \left( c + \frac{s_2}{r} \right) p_{22} - 2p_{12} - q_{22} + \frac{s_2^2}{r} = 0 \quad (4.7b)$$

$$p_{11} = \frac{1}{m^2 r} p_{12} p_{22} + \frac{1}{m} \left( c + \frac{s_2}{r} \right) p_{12} + \frac{1}{m} \left( k_1 + k_2 + \frac{s_1}{r} \right) p_{22} - q_{12} + \frac{s_1 s_2}{r} \quad (4.7c)$$

Controller:

$$u = - \frac{1}{r} \left( \frac{p_{12}}{m} + s_1 \right) y - \frac{1}{r} \left( \frac{p_{22}}{m} + s_2 \right) \dot{y} \quad (4.8)$$

By using the second-order method,  $p_{12}$  is computed first. It then substitutes  $p_{12}$  into Eqn (4.7b) to determine  $p_{22}$ . Eqn (4.7c) is a linear equation of  $p_{11}$  thus  $p_{11}$  can be obtained by direct substitutions. It can be seen that the number of nonlinear unknowns has been reduced which would improve the computational speed.

### **Case 3: Second-order velocity feedback**

Performance index:

$$J = E \left\{ \int_{t_0}^{t_f} (q_{22} \dot{y}^2 + 2s_2 \dot{y} u + r u^2) dt \right\} \quad (4.9)$$

Steady-state Riccati equation:

$$\frac{1}{m^2 r} p_{22}^2 + \frac{2}{m} \left( c + \frac{s_2}{r} \right) p_{22} - q_{22} + \frac{s_2^2}{r} = 0 \quad (4.10)$$

Controller:

$$u = - \frac{1}{r} \left( \frac{p_{22}}{m} + s_2 \right) \dot{y} \quad (4.11)$$

It is observed that the number of nonlinear unknowns for the velocity feedback system has been further reduced. It certainly will speed up the computation and data processing time.

## **Conclusions**

In this report, second-order equations were used to derive full-state and velocity feedback controllers. The Kalman filter for full-state and pure velocity measurements is also developed. The second-order method simplifies the process for controller and Kalman filter design. It allows the properties of second-order equations be preserved. The number of nonlinear unknowns can be reduced by nearly 25% for the full-state systems and nearly 75% for the velocity feedback systems. A full-state system requires sensors to measure both positions and velocities and then feeds these signals back to the actuator to generate the desired force. However, an optimal damping controller only measures velocities and these signals are fed back to determine the actuating force. With less number of parameters to be measured and computed, the damping control system will certainly improve the computation and data processing time. With less measurements imply that the hardware installations may be simplified. An example is used in this report to illustrate the procedure of how these controllers are determined. The example demonstrates that second-order methods hold a number of advantages over the first-order approach. Velocity feedback systems can simplify the design of controllers allowing the hardware and software installations be less expansive and more efficient.

## **References**

- Balas MJ, 1979, "Direct Velocity Feedback Control of Large Space Structures," *Journal of Guidance and Control*, Vol. 2, No. 3, pp. 252-253.
- Bryson AE and YC Ho, 1969, *Applied Optimal Control*, Blaisdell Pub: Waltham MA.
- Friedland B, 1988, *Control Systems Design: An Introduction to State-Space Methods*, McGraw-Hill, New York.
- Hashemipour HR and AJ Laub, 1988, "Kalman Filter for Second-Order Models," *Journal of Guidance and Control*, Vol. 11, pp. 181-186.
- Hamidi M and Juang JN, 1981, "Optimal Control and Controller Location for Distributed Parameter Elastic Systems," 20th IEEE Conference on Decision and Control, pp. 502-506.
- Kwak MK and L Meirovitch, 1990, "An Algorithm for the Computation of Optimal Control Gains for Second-Order Matrix Equations," AIAA-90-3503-CP.

Lieh J, 1992a, "Semiactive Damping Control of Vehicle Ride," 1992 ASME Winter Annual Meeting, Active Control of Noise and Vibration, DSC-Vol. 38, pp. 345-351.

Lieh J, 1992b, "Optimal velocity control of Second-Order Systems with Cross Product in the Performance Index," 1992 ASME Winter Annual Meeting, Active Control of Noise and Vibration, DSC-Vol. 38.

Skelton R, 1988, *Dynamic Systems Control: Linear Systems Analysis and Synthesis*, Wiley: New York.

Stengel RF, 1986, *Stochastic Optimal Control: Theory and Application*, Wiley: New York.

# **PARALLEL PROCESSING FOR REAL-TIME RULE-BASED DECISION AIDS**

**Chun-Shin Lin  
Department of Electrical and Computer Engineering  
University of Missouri-Columbia  
Columbia, MO 65211**

**Final Report for  
Summer Faculty Research Program  
Wright Laboratory**

**Sponsored by:  
Air Force Office of Scientific Research  
Bolling Air Force Base, Washington D.C.**

**and**

**Wright Laboratory  
Pilot/Vehicle Interface Section  
Wright-Patterson AFB, Dayton OH**

**August, 1995**

## PARALLEL PROCESSING FOR REAL-TIME RULE-BASED DECISION AIDS

Chun-Shin Lin  
Department of Electrical and Computer Engineering  
University of Missouri-Columbia

### Abstract

The rapid technology development in the past two decades has made today's combat a complicated task. A large amount of information can be available in a mission from both on-board and off-board sources. Effectively utilizing the information is necessary to achieve successful and optimal results. Decision aids that operate in real-time are an important issue as all DoD components strive to reduce the crew size of their various weapon systems. The decision aids will help reduce the workload of the crew and increase the efficiency and reliability of operations. Since a large number of criteria and rules must be evaluated and checked in a very short time period in combat automation, parallel processing may be needed in order to meet the timing requirement. In this project, we investigate a parallel processing technique for complicated decision aids that employ two-state rule-based systems. The study focuses on evaluation of decision rules, although that is just part of the decision aids problem. The rule base is decomposed into subsets for individual processing units. The rule-checking task is distributed to multiple processors to speed up the response. One merit of the explored technique is the scaleability. The number of processors can be altered based on the processing load and the availability of processors. The Intel Paragon high performance computer (a 2-dimensional mesh processor architecture) is selected for experiments. This report introduces the data structures for rule bases and the developed software that was used in this study and which can be used in the future. Experimental results using different numbers of processors are presented.

# **PARALLEL PROCESSING FOR REAL-TIME RULE-BASED DECISION AIDS**

Chun-Shin Lin

## **1. INTRODUCTION**

The rapid development of advanced technology in the past two decades has made today's combat a complicated task. A large amount of information is available during a mission from both on-board and off-board sources. Effectively utilizing the information is necessary to ensure successful and optimal results. Decision aids that operate in real-time are an important issue as all DoD components strive to reduce the crew size of their various weapon systems. The decision aids will help reduce the workload of the crew and increase the efficiency and reliability of operations.

Technologies for rule based systems, fuzzy logic and neural networks are all important to the design of an advanced decision-aid system. Although a lot of research efforts have been devoted into these areas in the recent years, the level of practical usage of these technologies on air fighters and other weapon systems is far below expectations. Insufficient computational power is often one of the factors that hold back the applications.

The problem of computing power may be solved by a hardware or software approach. Special hardware architectures or chips for expert systems and neural networks [1] have been suggested in many studies. The special-purpose hardware does provide fast computing speed with a tradeoff in less flexibility. Special hardware often becomes totally useless if it doesn't completely fit the requirements of an application. In such a situation, design and fabrication of a new chip, or implementation of a new board or a new system may become necessary. In other words, special hardware could cause drawbacks such as poor flexibility, long redesigning time, high cost, short life-time, etc. These drawbacks may be part of the reason commercially available neural network chips are not as popularly adopted as expected.

The software approach is an alternative attractive to designers of intelligent systems [2]. The flexibility makes the application design easier and is likely to help extend the life-time of the developed system due to relatively easy modification and upgrading. The software approach is becoming more attractive and feasible, while the general purpose high performance multiprocessor system is expected to



become more popular and affordable in the near future. The parallel processing technology for implementing intelligent systems on general-purpose multiprocessor architectures demands more attention.

In this study, we have explored the development of a scaleable intelligent decision aid that employs a two-state rule-base scheme. The structure consists of four major portions [3,4]: (1) information collection, (2) information processing and criteria evaluation, (3) rule checking, and (4) action execution. The information is collected by sensors. The information processing may involve conventional computation and algorithms, as well as neurocomputing and fuzzy logic. Rule-checking will take the criteria values (binary) and determine which rules should be fired. Actions that are associated with the fired rules will then be executed. Figure 1 shows a basic structure. Our long-term goal is to develop an efficient *scaleable* parallel processing technique for implementing such a structure. The design should be automatically reconfigurable based on the available hardware resource and the processing load.

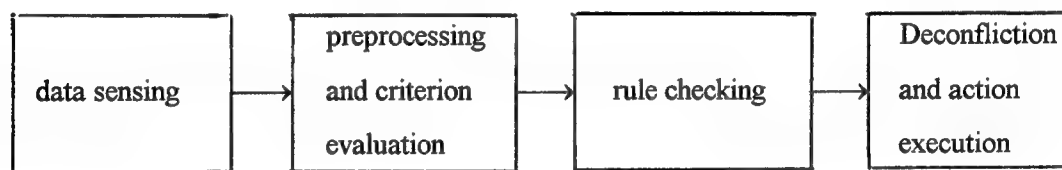


Figure 1. Basic components in a two-state rule-based decision aid

In the first 8 weeks of study on this problem, efforts have been focused on rule checking. Major accomplishments include the following:

- Suggested a structure for parallel processing in rule-based decision aids.
- Developed software tools for current and future experimental study.
- Provided experience and experimental results showing how the use of parallel processing will help handle large-size rule-based decision aids.

Note that this study only focuses on one part of the proposed structure. Further research is needed on other parts and system integration.

## 2. THE OVERALL INTELLIGENT DECISION AID

The basic block diagram of an intelligent decision-aid has been given in Figure 1. Information is collected by sensors. The sensed information may be preprocessed to generate the derived data for the criterion evaluator. One simple example of derived data is the rate of change of a sensed value. Neural networks may be used in preprocessing too. The outputs of the criterion evaluator are binary values (two-state). A criterion indicates whether a special condition or a sequence of conditions are satisfied or not. One example is that a specific voltage value has been kept over 5V in the past three time intervals. Criteria are inputs to the rule-checking module.

$C_i$  is used to denote the  $i$ th criterion, which has a value of either 0 or 1.  $\sim C_i$  denotes the complement of  $C_i$ . A rule is represented as a logic minterm (AND of Boolean variables) [3,4]. For instance,

$$R_k : (\text{action list}) \leftarrow C_2 \& \sim C_5 \& C_{12} \quad (1)$$

where "&" denotes the logic AND. The rule in (1) is fired when  $C_2 = 1$ ,  $C_5 = 0$  and  $C_{12} = 1$ . When the rule  $R_k$  is fired, actions in the action list will be executed. Displaying a piece of information to the pilot, recommending an action or even taking over part of a pilot's tasks are examples of actions.

## 3. PARALLEL PROCESSING SYSTEMS

This study on parallel processing assumes a 2-dimensional mesh processor architecture. Each processor (a node) can execute its own program and communicate with others through some SEND and RECEIVE commands. The Intel Paragon Computer [5] available in Wright Laboratory for defense research studies belongs to this type. This Paragon consists of 352 general-purpose nodes called GP nodes. Each GP node has a single i860 XP application processor, as well as an additional i860 XP as a message processor for message operations. When an application decides to send a message, the message processor handles the work and frees the application processor to continue with numerical computing. Each GP node has its own 32 Mbytes of memory. The computer system is scaleable and can be easily expanded by

adding new nodes. Since the computer is a multi-user system, interference between different processes may exist due to data transmission.

It is noted that the Intel i860 and i960 are used on today's military aircraft. Thus the results from the study using the Paragon are more easily transferable to practical use in operational environments.

With this kind of structure, one can decompose the rule-checking task evenly into  $p$  processors and have each processor evaluate the assigned rules. The structure will be easy to reconfigure due to the simple decomposition. The selection of  $p$  should be based on the availability of nodes and the processing load.

#### 4. KNOWLEDGE REPRESENTATION IN SCALEABLE TWO-STATE RULE-BASED SYSTEMS

As introduced earlier, the rule-based system will have the rules represented in the form

$$R_k : (\text{action list}) \leftarrow C_i \dots \& \sim C_j \dots \& C_m$$

The data structure must indicate which criteria are included in each rule. The data structure is illustrated in Figure 2. A list of criteria, called LIST\_CRITERIA, used by all rules is constructed. A number  $i$  in the list indicates that  $C_i$  is included and  $-i$  indicates that  $\sim C_i$  is included. Another list RULE\_POINTER stores the positions of the last criteria of all rules (see Figure 2) [3-4]. For example, RULE\_POINTER[j] is the pointer to the last criterion used by the rule  $R_j$ . If  $a = \text{RULE\_POINTER}[j-1]$  and  $b = \text{RULE\_POINTER}[j]$ , then the rule  $R_j$  uses the criteria denoted by LIST\_CRITERIA[a+1], LIST\_CRITERIA[a+2], . . . . LIST\_CRITERIA[b]. Note that  $h$ ,  $p$  and  $q$  in the figure are positive or negative criteria numbers.

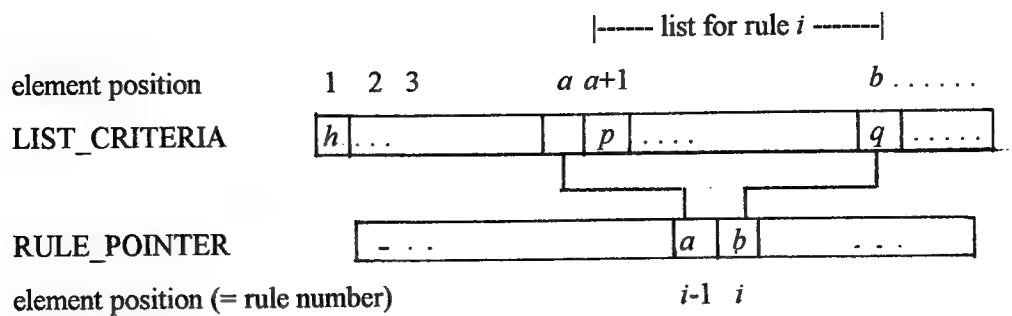


Figure 2. Data structure for rule base.

The system will have a rule fired at the time when all involved criteria become true. The rule will be kept at a fired status until one or more involved criteria become unsatisfied. At any time, only the rules involving the criteria that change values need to be checked. Thus it is more efficient to construct a data structure to make it easier to find the set of rules that need to be checked. This means backward pointers from criteria to rules are needed. This requires the construction of a data structure similar to the one above. Figure 3 shows such an index structure for backward pointers. CRITERION\_POINTER provides information for quickly determining which subset of rules in the list LIST\_RULES should be evaluated.

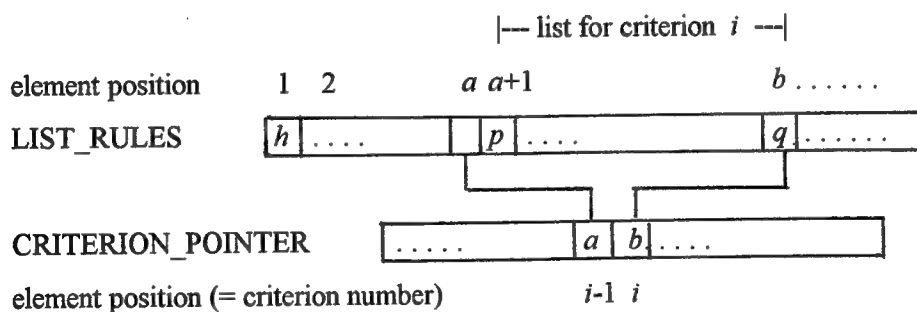


Figure 3. An index structure denoting which rules are used by a criterion

Note that the first data structure consists of complete knowledge and the second one can be derived from it.

### Rule-Base Knowledge for Each Processing Node

Since each node will check only a subset of rules, it doesn't need the complete rule-base knowledge. The data structure for the subset of rules can be represented in a similar structure as that for the overall knowledge shown in Figures 2 and 3. However, only a sublist from the LIST\_CRITERIA and a sublist from RULE\_POINTER will be stored for each processor. In the index data structure, the length of CRITERION\_POINTER will remain the same but the rules in RULE\_LIST not handled by the assigned node will be removed. The subsets of rule bases can be generated from the overall rule base by a computer program.

## **6. RULE EVALUATION AND SUPPORTING TOOLS**

Figure 4 shows the general structure of the rule evaluation and supporting tools. The overall rule base should be generated and be modified through rule generation and modification utility software. According to the assigned number of processors, the knowledge base should be decomposed into knowledge bases for individual nodes. This must be done by a knowledge decomposition program. Each node of the multiprocessor rule-checking system will then read in its own knowledge base. Although the knowledge decomposition program and the rule-checking program have been separated in this work, they can be easily combined. The rule-checking program will receive the information regarding the criteria values from the criterion evaluation program. The data passed to the rule-checking program is a list of criteria numbers indicating which criteria change values in the most recent time interval.

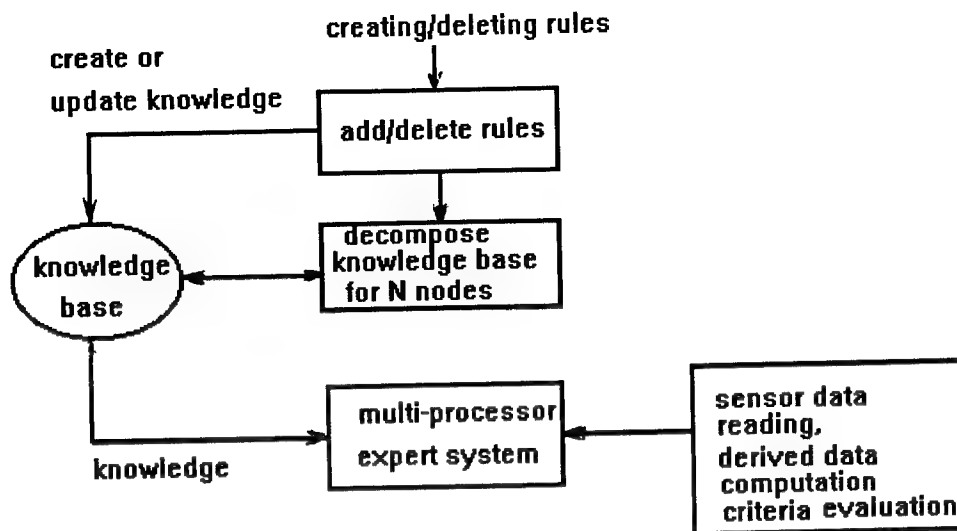


Figure 4. Rule checking and supporting software

## 7. SOFTWARE FOR EXPERIMENTS

With the suggested data structure, knowledge representation, and the decomposition of the rule-checking task, we are interested in evaluating how the use of parallel processing will help. Major programs have been developed for the experiments:

- Rule generation program: RULEGEN

For experimental purposes, we have modified one of Noyes' programs (developed for the work described in [4]) to generate rules in a random way. The program generates a set of rules with a structure shown in Figure 2.

- Index generation program: INDXGEN

This program generates the data structure shown in Figure 3. Using the generated data structure, one can easily find which rules involve a specific criterion.

- Rule decomposition program: R\_DECOMP

This program reads in the number of assigned nodes and decomposes the knowledge base into subsets for individual nodes. The program developed in this study can be run on either the Paragon or

a SUN workstation. It is possible to combine this program with the rule-checking program RCHECK below.

- Rule checking program: RCHECK

This program is for the Paragon computer. Node 0 simulates the criterion evaluator and all others perform rule checking. At the beginning, each node will read in its subset of the rule base from a file generated by R\_DECOMP and read in initial criterion data from the criterion evaluator (node 0). The rule-checking node will then receive a list of changed criteria broadcast from the criterion evaluator in each processing cycle. The involved rules will then be checked and the fired rules identified. The following pseudo code describes the program structure:

```
if ( mynode( ) == 0)
{ global synchronization;
  generate the list of changed criteria;
  broadcast the list of changed criteria;
}
if ( mynode( ) > 0)
{ global synchronization;
  receive the list of changed criteria;
  check the assigned rules that involve changed criteria;
}
```

Some synchronization is needed in this program. The rule-checking nodes will perform rule checking only when a new criterion-change-list is received. Node 0 will send out new data whenever all other nodes have completed the previous rule-checking task. Details on the use of the software are provided in Appendix A.

## 8. EXPERIMENTS AND RESULTS

Four sets of rule bases have been generated by RULEGEN and INDXGEN for experimental tests. Each set is named with five digits. The first digit specifies the number of rules, the second and third denote the number of different criteria in the system, and the last two digits indicate the maximum number of criteria in a rule. The four test sets are

- (1) Set 31508: 3000 rules, 1500 different criteria, each rule can have up to 8 criteria (an average of 4).
- (2) Set 44010: 4000 rules, 4000 different criteria, each rule can have up to 10 criteria (an average of 5).
- (3) Set 54010: 5000 rules, 4000 different criteria, each rule can have up to 10 criteria (an average of 5).
- (4) Set 64010: 6000 rules, 4000 different criteria, each rule can have up to 10 criteria (an average of 5).

The criterion-evaluation simulator generates a list of changed criteria. Different problems could have very different percentage of criteria that change values in each evaluation period. In this study, the maximum size of the list was selected to be 75 for the test set 31508, and 200 for others (*i.e.*, at most 5% of criteria change value in each time interval). The size and the criteria numbers were generated by a random number generator in the experiments. The size affects the time of data transfer and the number of rules to be checked. For each test set, we ran the scaleable program on 1, 2, 3, 4, 5, 6, 7, 10, and 15 nodes. Note that one additional node was required for criterion evaluation. Each program was run for 200 cycles with each cycle including the generation of a new list of changed criteria and the evaluation of involved rules. Three different kinds of computer times were collected. They are

1. the total time used in receiving the list of changed criteria from the criterion-evaluation simulator,
2. the total time spent on rule checking, and
3. the total time for completing 200 cycles.

The third one includes the waiting time for synchronization. As mentioned previously, the Paragon is a multi-user system. The operating system assigns node partition and data of other processes may be transmitted through the allocated nodes. These cause the timing data collected in our experiments to be slightly different for different runs. Timing data for the four sets are provided in Tables 1-4. For cases with 10 and 15 nodes, there is a possibility that the timing data are unreasonably large (such as 20



seconds). The operating system may have removed the programs and reallocate nodes later. Those data were discarded.

The largest data transmission time, the longest rule-checking time and the total processing time for the set 64010 are plotted in Figure 5. Plots for other three all look like the one shown in Figure 5 and are not included.

Table 1. Computer times for test set 31508

# of nodes	range of time for receiving data	range of time for evaluating rules	total time for completing 200 cycles
1	0.0157	0.8561	0.8933
2	0.0325-0.0328	0.4075-0.4148	0.4804
3	0.0241-0.0338	0.2743-0.2758	0.3580
4	0.0217-0.0385	0.2054-0.2124	0.3046
5	0.0220-0.0461	0.1549-0.1564	0.2586
6	0.0286-0.0455	0.1251-0.1334	0.2341
7	0.0232-0.0516	0.1103-0.1140	0.2238
10	0.0246-0.0623	0.0786-0.0839	0.2078
15	0.0279-0.0694	0.0542-0.0588	0.2017

(The unit for time is second.)

Table 2. Computer times for test set 44010

# of nodes	range of time for receiving data	range of time for evaluating rules	total time for completing 200 cycles
1	0.0167	1.1597	1.1980
2	0.0357-0.0369	0.5748-0.5767	0.6542
3	0.0258-0.0368	0.3936-0.3975	0.4878
4	0.0237-0.0426	0.3035-0.3089	0.4166
5	0.0234-0.0502	0.2513-0.2588	0.3894
6	0.0324-0.0509	0.2029-0.2588	0.3424
7	0.0259-0.0572	0.1776-0.1812	0.3238
10	0.0285-0.0672	0.1312-0.1363	0.3095
15	0.0324-0.0785	0.09757-0.1018	0.2996

(The unit for time is second.)

Table 3. Computer times for test set 54010

# of nodes	range of time for receiving data	range of time for evaluating rules	total time for completing 200 cycles
1	0.01682	1.4487	1.4880
2	0.0341-0.0349	0.7244-0.7259	0.8027
3	0.0261-0.0367	0.4849-0.4896	0.5850
4	0.0233-0.0430	0.3708-0.3790	0.4852
5	0.0239-0.0510	0.3032-0.3151	0.4416
6	0.0311-0.0511	0.2628-0.2698	0.4404
7	0.0324-0.0504	0.2155-0.2214	0.3564
10	0.0279-0.0670	0.1574-0.1659	0.3348
15	0.0317-0.0785	0.1131-0.1202	0.3173

(The unit for time is second.)

Table 4. Computer times for test set 64010

# of nodes	range of time for receiving data	range of time for evaluating rules	total time for completing 200 cycles
1	0.0173	1.7385	1.7780
2	0.0348-0.0364	0.8641-0.8730	0.9492
3	0.0265-0.0378	0.5786-0.5828	0.6705
4	0.0247-0.0428	0.4377-0.4458	0.5504
5	0.0242-0.0517	0.3576-0.3645	0.4948
6	0.0324-0.0506	0.3050-0.3116	0.4607
7	0.0257-0.0580	0.2686-0.2756	0.4192
10	0.0278-0.0677	0.1824-0.1914	0.3579
15	0.0317-0.0762	0.1298-0.1377	0.3266

(The unit for time is second.)

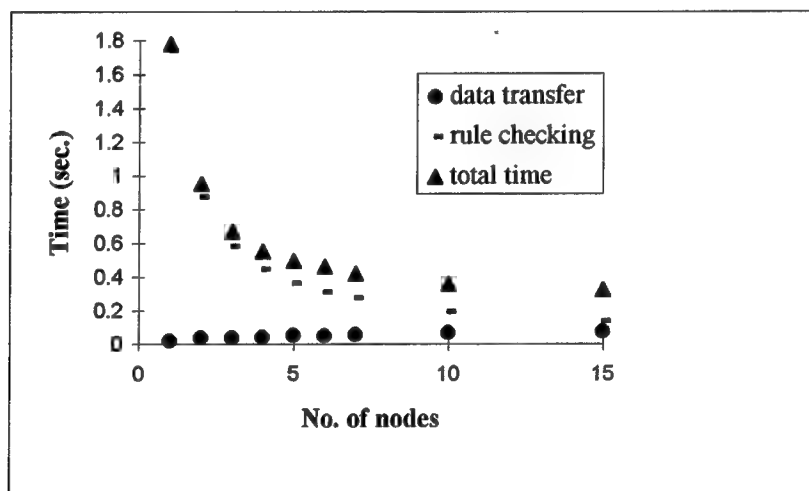


Figure 5. Timing data for the set 64010 (data from Table 4)

Some observations in Tables 1-4 and Figure 5 are discussed below:

- The rule-checking time is approximately inversely-proportional to the number of nodes. This result was expected.
- The total processing time for 200 cycles is not equal to and may not be close to the sum of the maximum data transmission time and the rule-checking time. The difference is caused by the node idle time. Although the number of rules assigned to each node is the same or about the same, the processing time periods on different nodes for each cycle are different. Balancing the number of assigned rules does not balance the processing (rule-checking) load in each cycle. The data show that the time wasted on waiting (estimated by subtracting the maximum data transmission time and the maximum rule-checking time from the total processing time for 200 cycles) becomes more significant while more nodes are used. When the time for rule-checking is close to the sum of the waiting time and the data transmission time, the parallelism becomes less efficient. In these test runs, it happened at around 6-10 nodes.
- The total time for the criterion-evaluation simulator to generate two hundred lists of changed criteria (not including data passing) is about 0.054 sec. for 31508 and 0.143 sec. for 64010. This time period is also a factor that should be considered in determining the number of nodes to be used. There could not be further improvement by adding more nodes if the rule-checking time has been already smaller than the criterion-evaluation time. For all tested cases, this happens at about 15 nodes.
- Data transmission time for 31508 is slightly smaller than that for 64010 because of the smaller size of data transferred. However, the difference is small. This indicates that most time is used to set up the data communication and a small amount of time is actually used for sending data. A set of 0~200 integers is considered a small record.
- With more nodes, the data transmission time is longer.

To measure the efficiency of parallelism, we define the speed-up measurement as

$$\text{Speed-up} = \text{time to execute on one processor} / \text{time to execute on } p \text{ processors}$$

where the time is the total time for completing 200 cycles as given in the tables. The “*speed-up*” value for the cases 44010, 54010 and 64010 are plotted in Figure 6. The only difference in these three cases is the number of rules. However, the test set 31508 has all the number of rules, the number of criteria and the maximum number of criteria in each rule different. It is not meaningful to compare its speed-up value with the other three. Thus the speed-up curve for 31508 is not included in this Figure. Among the three compared, the set 64010 has the largest rule base. From the plot, it is seen that the increase of the speed-up value gets slower after around 4 nodes for these sizes of rule-base systems. This change is mainly due to the increase of data transmission time and the waiting time, and the decrease of the rule-checking time. If the load can be better balanced and the waiting time can be reduced, one will be able to use more nodes without losing the efficiency of parallelism. More discussion on this problem is given on the next section. It is also seen from the figure that the speed-up value for the larger system is greater.

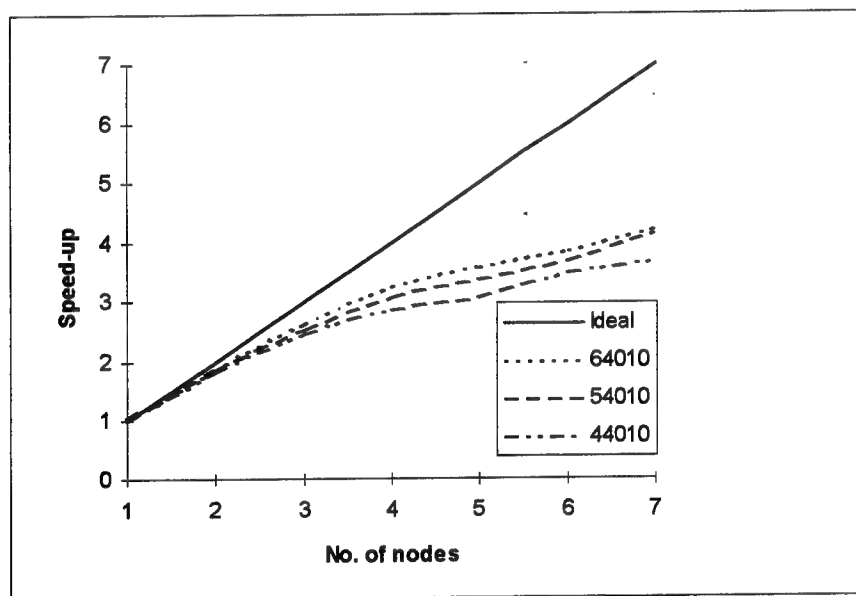


Figure 5. Speed-up factor

## 9. DISCUSSION AND CONCLUSIONS

Decision aids that operate in real-time are an important issue as all DoD components strive to reduce the crew size of their various weapon systems. To process a large amount of available information in real-time, parallel processing seems to be one solution. In this project, we investigated the parallel processing technique for complicated decision aids that employ two-state rule-based systems. The study focuses on evaluation of decision rules -- part of the decision aids problem. The rule base is decomposed into subsets for individual processing units. The rule-checking task is distributed to multiple processors to speed up the response. One merit of the explored technique is the scalability. The number of processors can be altered based on the processing load and the availability of processors. The Intel Paragon high performance computer (a 2-dimensional mesh processor architecture) was selected for experiments

Four sets of rule bases have been generated for experimental tests. For each test set, the scaleable program was run on 1, 2, 3, 4, 5, 6, 7, 10 and 15 nodes. If it is required that rules be checked in an average time period of 2.5ms, from the data given in Table 4, one would suggest that five processors be used for a rule-based system like "64010" (6000 rules, 4000 criteria, at most 10 criteria in each rule, and at most 200 criteria changing values in each cycle). It is seen that the total time for five processors in Table 4 is 0.4948 sec. and the average time period,  $0.4948/200$  sec., is shorter than 2.5ms. It is also noted that although the number of rules assigned to each node is the same or about the same, the processing time periods on different nodes for each cycle are different. This causes the waiting (idle) time. The waiting time changes from time to time depending on which criteria change values. The waiting time of each node is a non-deterministic value. Balancing the number of assigned rules does not balance the processing load in each cycle.

The scheme investigated in this study decomposes the rule base into subsets and assigns one node to check a subset of rules. As seen from the experimental results, the load is not perfectly balanced due to the non-deterministic waiting time. The load balancing may be improved by decomposing the rule base into smaller subsets and assigning each subset to at least two processors. Although a subset is assigned to two or more nodes, it will be checked by only one processor, the one that becomes available first. With such an arrangement, a processor may handle more subsets with lighter processing load or less subsets with heavier processing load. This should improve the load balancing and efficiency of parallelism. Another important

merit is improved fault-tolerance. Failure of nodes that are not assigned common subsets will not fail the system. Further study should be devoted to investigating this new idea.

Efforts should also be devoted to other components of this kind of rule-based decision aid. Neural networks and the criterion evaluator are two major types of components. A similar idea introduced above can be applied to parallelizing the neurocomputing. Neurons in the same layer or set of basis function units may be grouped into subsets and assigned to different processors. Assigning each subset to at least two processors will also create the fault tolerance. One difference between the neurocomputing and rule-checking is that the processing time for the former one is deterministic.

Another issue is system integration. It has been seen that some processes have deterministic processing time and others have non-deterministic time. In addition these processes have some precedence relationship. With the different properties and precedence relationship, how can one best assign these processes to the available nodes is one problem requiring a study. An automatic process-assignment program is desirable if the scalability and reconfigurability are desired.

## **REFERENCES**

1. Special Issue on Neural Network Hardware. IEEE Transactions on Neural Networks, May 1992, Vol. 3, No. 3, pp 347-506.
2. Kanal, L. N., et al. (Ed), Parallel Processing for Artificial Intelligence, Vol. 1 and 2, North-Holland, 1994.
3. Raeth, P. G., J L. Noyes and A. J. Montecalvo, "A scaleable frame work for adding crisp knowledge to pilot/vehicle interfaces," 1994 IEEE International Conference on Systems, Man, and Cybernetics, pp. 2091-2096, Oct. 1994.
4. Noyes, J. L., "Expert system rule-base evaluation using real-time parallel processing," WL-TR-93-3098, Fight Dynamics Directorate, Wright Laboratory, Wright Patterson AFB, Ohio.
5. Paragon User's Guide. Intel Corporation, June 1994.

## **ACKNOWLEDGMENTS**

I would like to express my great appreciation to Major Peter G. Raeth, Chief of the Pilot/Vehicle Interface Section. He made the arrangement to have me attend the Paragon Programming course, which is important to this project. He provided me related literature that helps me better understand the problem. Many

valuable discussions with him are the major key for the successful completion of this project in the short eight-week period. I would also like to thank the Air Force Office of Scientific Research for sponsoring this study through the Summer Faculty Research Program.

## **APPENDIX: USAGE OF THE EXPERIMENTAL SOFTWARE**

The following is the step by step procedure in using the experimental software:

### **1. Generate simulation rules**

- Run RULEGEN.PAS under turbo PASCAL on a PC and provide the number of rules, task names, etc.
- There will be three output files: r1\_(task name).txt, r2\_(task name).txt and r3\_(task name).txt, where task name is the one provided by the user while running the program.. r1\_(task name) will be used later by another program r\_decomp.c. r3\_(task name).txt will be sorted for creating another needed file for r\_decomp.c. r2\_(task name).txt is a list of rules for user's reference and will not be used by any other program.

### **2. Sorting**

- Send r3\_(task name).txt to a UNIX system. In this work, the FTP command and a SUN workstation were used.
- Use the following command to sort the file: sort +1 -2 -o s3\_(task name).txt r3\_(task name).txt. It is preferred that the same task name is used all through the work. The two-column file will be sorted according to the second column. The new file provides the explicit information about which rules are used by a specific criterion.
- Get the file s3\_(task name) back to PC.

3. Check the number of lines in the file s3\_(task name).txt and add the number onto the top of the file.

4. Generate the criterion index file x2\_(task name).txt.



- Run INDXGEN.PAS under turbo PASCAL on a PC and provide the task names.
  - Two output files will be generated: x1\_(task name).txt and x2\_(task name).txt. The file x2\_(task name).txt is used by r\_decomp.c. x1\_(task name).txt is a list of rules according to criteria.
5. Send x2\_(task name).txt to sd1.wpafb.af.mil for PARAGON use.
  6. On sd1.wpafb.af.mil or paragon, compile r\_decomp.c and run it. Give the number of processors to be used for rule processing. R3\_(task name).TXT and X2\_(task name).TXT are used by this program.. The output will be KNOW1, KNOW2, ..., KNOWn, where n is the number of processors. These subfiles are knowledge bases for individual processors.
  7. Rule processing
    - Compile rcheck.c by "icc -nx rcheck.c" on sd1 or paragon.
    - Run the program on (n+1) nodes. Timing information will be provided on the screen and the fired rules at every time instant can be listed into "fire#.txt" if the print instruction at the end of the program is enabled.

**A STUDY OF DEAD RECKONING  
AND SMOOTHING IN  
DISTRIBUTED INTERACTIVE SIMULATION**

**Kuo-Chi Lin  
Associate Professor  
Institute for Simulation and Training  
University of Central Florida  
3280 Progress Drive  
Orlando, FL 32826-0544**

**Final Report for:  
Summer Faculty Research Program  
Wright Laboratory**

**Sponsored by:  
Air Force Office of Scientific Research  
Bolling Air Force Base, DC  
and  
Wright Laboratory**

**August 1995**

**A STUDY OF DEAD RECKONING  
AND SMOOTHING IN  
DISTRIBUTED INTERACTIVE SIMULATION**

**Kuo-Chi Lin  
Associate Professor  
Institute for Simulation and Training  
University of Central Florida**

**ABSTRACT**

The main objective of the Summer Research Faculty is to make sure the ITB Simulator at Wright Laboratory is DIS compatible in dead-reckoning algorithm. This report analyzes the use of dead reckoning in Distributed Interactive Simulation. The purpose of dead reckoning is to reduce the updates required by each simulator on the network to better utilize the available bandwidth. Extrapolation formulas are derived and discussed. Smoothing and delay compensation algorithms are also discussed. A software tool that assesses the performance of the read reckoning algorithm is introduced. After implementing the body-axes dead-reckoning algorithms, the goal of the Summer Research Faculty is achieved.

# A STUDY OF DEAD RECKONING AND SMOOTHING IN DISTRIBUTED INTERACTIVE SIMULATION

Kuo-Chi Lin

## 1. INTRODUCTION

The main objective of the Summer Research Faculty is to make sure the ITB Simulator at Wright Laboratory is DIS compatible in dead-reckoning algorithm. Distributed Interactive Simulation (DIS) is an exercise involving the interconnection of a number of simulators in which the simulated entities are able to interact within a computer generated environment.<sup>1</sup> The simulators may be present in one location or be distributed geographically. The communication between simulators is provided by a computer network developed under an advanced research project sponsored by the Defense Modeling and Simulation Office (DMSO) in partnership with the Simulation Training and Instrumentation Command (STRICOM).<sup>2,3</sup>

As a simulator models the behavior of a vehicle in real time, that vehicle's position/orientation is constantly changing. The vehicle's simulator must inform other simulators of these changes so that all simulators participating in the exercise can depict the vehicle correctly, at its current location.

To reduce the communication traffic in the network, the position/orientation of each entity will not be sent through the network every single time a change occurs. Instead, a technique call "dead-reckoning" is used. The term, borrowed from navigation, means establishing the position of a ship based on an earlier known position and estimates of time and motion. Simulator may use dead reckoning to extrapolate the position/orientation of vehicles thus reducing the frequency of which the simulators would have to obtain the actual information from the network.

## 2. DEAD RECKONING

Dead reckoning is used in the following manner.<sup>4</sup> Each simulator is responsible for maintaining a detailed model of its own vehicle's state and a precise notion of its own position/orientation over time. In addition, each simulator also maintains a simple dead reckoning model of the state of all other vehicles with which it might possibly interact. The dead reckoning model is maintained until such time as a new information package is received from the network.

This approach implies that each simulator is also responsible for issuing a new information package of its own appearance whenever it is necessary. To do this each simulator must maintain in addition to its high fidelity model a dead reckoning model that corresponds to the model that other simulators are maintaining of its vehicle. After each update of its high fidelity model and its dead reckoning model, the simulator compares the exact appearance of its vehicle with the extrapolated appearance and issues a new information package to the network only when the error exceeds a threshold. Figure 1 illustrates this concept. At time  $t_n - \Delta$ , the vehicle detects that at the next update, time  $t_n$ , the error will exceed the threshold. Therefore, it sends out an information package immediately so that a new dead reckoning model can start at time  $t_n$ .

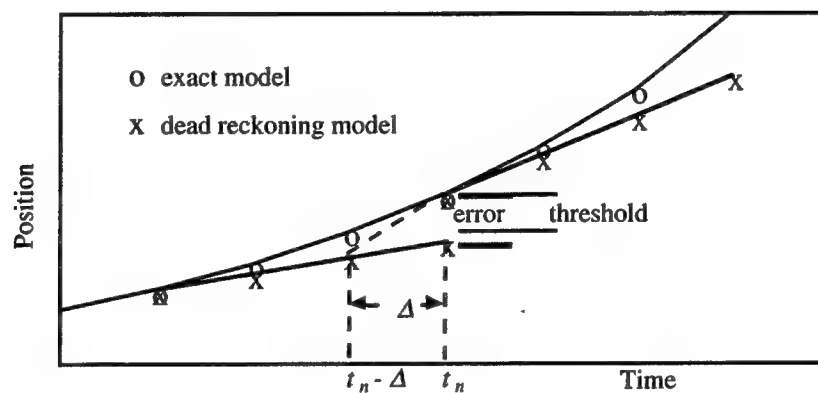


Figure 1. The concept of dead reckoning.

In essence, dead reckoning achieves a trade off among four factors: the network communication traffic, the amount of computation performed by simulators, and the precision and smoothness with which each simulator perceives of other simulators.

### 3. EXTRAPOLATION FORMULAS

The extrapolation formulas used in dead reckoning can be derived using the Taylor series expansion.<sup>5</sup> For simplicity, the one dimensional linear motion is used as an example to demonstrate different extrapolation formulas. Assume that  $a$ ,  $v$ , and  $x$  represent the acceleration, velocity, and position of the vehicle, respectively. The extrapolation formulas of different orders are given by:

$$\text{Zero Order} \quad x_i = x_0 \quad (1)$$

$$\text{First Order} \quad x_i = x_0 + v_0 \Delta \quad (2)$$

$$\text{Second Order} \quad x_i = x_0 + v_0 \Delta + \frac{a_0 \Delta^2}{2} \quad (3)$$

where the subscripts  $t$  and  $0$  represent the variables at time  $t = t_0 + \Delta$  and  $t_0$ , respectively, i.e., the position of the vehicle at time  $t$  is extrapolated based on the information received at time  $t_0$ . In general, using extrapolation formula of higher order can reduce the updates required by each simulator on the network; hence it causes less network traffic than using lower order formulas. On the other hand, higher order formulas need more information in each package sent through the network, thus creating a higher computation load for each simulator than the lower order ones.

The above formulas are one-step methods since they depend on the information at one point of the time only. Methods using information at more than one point of time are called multi-step methods. Some examples of the multi-step methods are given by:

$$\text{First Order} \quad x_i = x_0 + \frac{x_0 - x_{-1}}{T_{-1}} \Delta \quad (4)$$

$$\text{Second Order} \quad x_i = x_0 + v_0 \Delta + \frac{v_0 - v_{-1}}{T_{-1}} \frac{\Delta^2}{2} \quad (5)$$

$$x_i = x_0 - (x_0 - x_{-1}) \frac{\Delta^2}{T_{-1}^2} + v_0 \left( \Delta + \frac{\Delta^2}{T_{-1}} \right) \quad (6)$$

$$x_i = x_0 + \left( \frac{\Delta}{T_{-1}} + x_0 \frac{\Delta^2 + \Delta T_{-1}}{T_{-1}^2 + T_{-1} T_{-2}} \right) - x_{-1} \frac{(T_{-1} + T_{-2}) \Delta + \Delta^2}{T_{-1} T_{-2}} + x_{-2} \frac{\Delta^2 + \Delta T_{-1}}{T_{-2} (T_{-1} + T_{-2})} \quad (7)$$

$$\text{Third Order} \quad x_i = x_0 + v_0 \Delta + \frac{a_0 \Delta^2}{2} + \frac{a_0 - a_{-1}}{T_{-1}} \frac{\Delta^3}{6} \quad (8)$$

$$x_i = x_0 + v_0 \Delta + \frac{a_0 \Delta^2}{2} + \left( \frac{v_{-1} - v_0}{T_{-1}} + a_0 \right) \frac{\Delta^3}{3 T_{-1}} \quad (9)$$

$$x_i = x_0 + v_0(\Delta + \frac{\Delta^2}{T_{-1}}) + (x_{-1} - x_0)(\frac{3\Delta^2}{T_{-1}^2} + \frac{2\Delta^3}{T_{-1}^3}) + (v_0 + v_{-1})(\frac{\Delta^2}{T_{-1}} + \frac{\Delta^3}{T_{-1}^2}) \quad (10)$$

$$x_i = x_0 + v_0\Delta + v_0 \frac{3(2T_{-1} + T_{-2})\Delta^2 + 2\Delta^3}{6T_{-1}(T_{-1} + T_{-2})} - v_{-1} \frac{3(T_{-1} + T_{-2})\Delta^2 + 2\Delta^3}{6T_{-1}T_{-2}} + v_{-2} \frac{3T_{-1}\Delta^2 + 2\Delta^3}{6T_{-2}(T_{-1} + T_{-2})} \quad (11)$$

$$x_i = x_0 + v_0\Delta + (x_0 - x_{-1})\frac{\Delta^3}{T_{-1}^3} - v_{-1}\frac{\Delta^3}{T_{-1}^2} + a_0(\frac{\Delta^2}{2} + \frac{\Delta^3}{2T_{-1}}) \quad (12)$$

Here  $x_{-1}$ ,  $v_{-1}$ , and  $a_{-1}$  are displacement, velocity, and acceleration, respectively, of the vehicle updated at  $t_{-1} = t_0 - T_{-1}$ , one update before  $t_0$ ;  $x_{-2}$ ,  $v_{-2}$ , and  $a_{-2}$  are displacement, velocity, and acceleration, respectively, of the vehicle updated at  $t_{-2} = t_0 - T_{-2}$ , two updates before  $t_0$ . The multi-step methods need fewer data from the network than the one-step methods of same order. The tradeoffs are lower accuracy and higher memory requirements, which will be discussed in the later sections of this report.

In the current version of DIS standard, the entity acceleration and velocity are always sent together with the position in the Entity-State PDU. Therefore, using the multi-step extrapolation methods will not reduce the network traffic. In the future, if there is an attempt to reduce the size of the next-generation PDU, we can consider to use the multi-step extrapolation methods. In this case, the entity acceleration or even velocity need not be included.

There are other dead reckoning methods for some special situations. For example, assume that an aircraft is cruising at a constant altitude with periodic fluctuations. If the amplitude of the altitude fluctuation is within the threshold, the zero order dead reckoning may be better than the first order one. An example is shown in Figure 2. If the amplitude of the altitude fluctuation is greater than the threshold, a periodic term can be added to the zero order dead reckoning to represent the fluctuation. This additional term can reduce the network traffic drastically. This principle can be applied to the cases such as a ship on a wavy ocean as well as the missiles using a cyclic motion to sweep their tracking sensors.

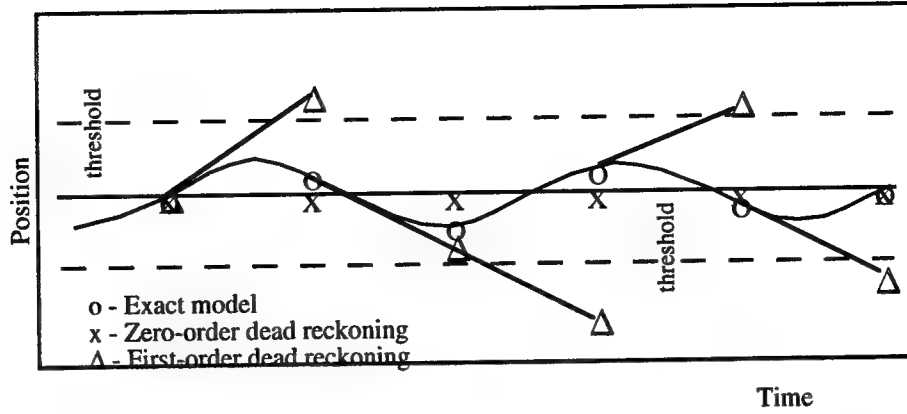


Figure 2. Example of periodic motion dead reckoning.

#### 4. SMOOTHING

When a new update of position is received from another entity, a correction in position is usually required so that the entity is depicted in simulation as correctly as possible. If the new position is put in the image display system immediately, it can cause jumps in the displayed image. Hence, the preferred method is to gradually correct the position of the entity over a period of time. This is called smoothing.

Figure 3 is an example of the smoothing technique. When the updated position is received at time  $t_n$ , instead of jumping to the new position, a smoothing model is maintained. First the new dead reckoning model is extrapolated using the updated information at time  $t_n$  to find the position at a future time, say  $t_n + 2\Delta$ ; then the positions at time  $t_n$  and  $t_n + \Delta$  is obtained by interpolating the positions at  $t_n - \Delta$  and  $t_n + 2\Delta$ . In this model, the transition from one dead reckoning model to another is smoother. One parameter is the number smoothing point. In the example shown in Figure 3, the number of smoothing points is two. In general, the number used is between five and 15.

There are several possible formulas for smoothing. Two examples are given by:

$$x_i = x_0 + (x_f - x_0) \frac{i}{(p + 1)} , \quad i = 1, \dots, p \quad (13)$$

$$x_i = [(p + 1)\Delta(v_0 + v_f) + 2(x_0 - x_f)] \frac{i^3}{(p + 1)^3} + [-(p + 1)\Delta(2v_0 + v_f) + 3(x_f - x_0)] \frac{i^2}{(p + 1)^2} + i\Delta v_0 + x_0 , \quad i = 1, \dots, p \quad (14)$$

where



- $x_i$ : the  $i$ -th smoothing position,  
 $i$ : integer, from 1 to  $p$ ,  
 $p$ : integer, number of smoothing points, (5, 10, or 15 in this experiment)  
 $x_0$ : starting position of smoothing, i.e., the position before update,  
 $x_f$ : final position of smoothing,  
 dead reckoning position  $p$  points after update.  
 $\Delta$ : the dead-reckoning step time,  
 $v_0$ : the velocity of the previous dead reckoning,  
 $v_f$ : the velocity of the next dead reckoning.

Here Eq. (13) is a straight-line equation and Eq. (14) is a cubic spline.

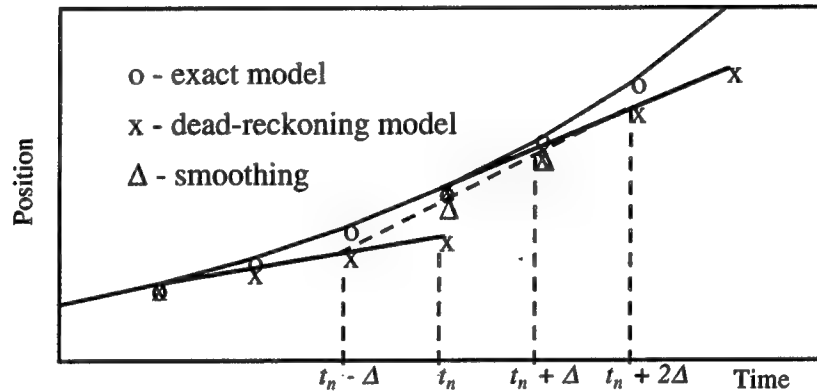


Figure 3. Example of smoothing technique.

## 5. SMOOTHNESS MEASURE

The smoothness of the visual display is a subjective measure. To obtain an objective measure, the variable that represents the motion of an entity on the screen has to be defined.<sup>6</sup>

The visual display of a simulator usually shows an image of the target seen by the pilot. At time  $t$ , the coordinates of the pilot are  $(x_p, y_p, z_p)$ , and the coordinates of the target are  $(x_g, y_g, z_g)$ . The relative position vector is given by

$$\mathbf{r} = (x_g - x_p)\mathbf{i} + (y_g - y_p)\mathbf{j} + (z_g - z_p)\mathbf{k} \quad (15)$$

where  $i, j$ , and  $k$  are unit vectors of the coordinate system. At time  $t + \Delta t$ , the relative motion between the target and pilot is

$$\Delta \mathbf{r} = (\Delta x_g - \Delta x_p)\mathbf{i} + (\Delta y_g - \Delta y_p)\mathbf{j} + (\Delta z_g - \Delta z_p)\mathbf{k} \quad (16)$$

where  $\Delta x$ 's,  $\Delta y$ 's, and  $\Delta z$ 's represent the coordinate changes in  $\Delta t$ . The linear motion appears on the screen is the component of the motion perpendicular to the line of sight, i.e.,

$$\Delta d = |\Delta \mathbf{r}| \sin[\cos^{-1}[\frac{\mathbf{r} \cdot \Delta \mathbf{r}}{|\mathbf{r}| |\Delta \mathbf{r}|}]] \quad (17)$$

The change of view angle in  $\Delta t$  is given by

$$\Delta \theta = \tan^{-1} \frac{\Delta d}{|\mathbf{r}|} \quad (18)$$

in radians. The geometric relationship is shown in Figure 4.

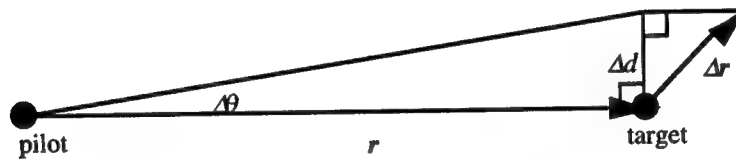


Figure 4. Relative motion between eye-point and target aircrafts.

The trajectory image smoothness can be measured by a formula in terms of  $\Delta \theta$ . One example is given by:

$$S = 1 - \frac{\sum_n \Delta \theta_{n+1} - \Delta \theta_n}{\Theta} \quad (19)$$

In this formulas, only the values of  $(\Delta \theta_{n+1} - \Delta \theta_n)$  that exceed the threshold are summed up. The smoothness,  $0 \leq S \leq 1$ , is a non-dimensional value normalized by the maximum value of  $\sum(\Delta \theta_{n+1}$

-  $\Delta\theta_n$ ),  $\Theta$ . The higher the value of  $S$  is, the smoother the image. Equation (19) is referred as " $S$  rating" in this report.

## 6. ORIENTATION DEAD RECKONING

The orientation dead reckoning equations, proposed by Burchfiel<sup>1,7</sup>, is a rotation matrix method based on the concept of Quaternions. This formula assumes the entity moves in a uniform manner (i.e.; constant rate turns,  $\omega = \text{constant}$ ), and is given by

$$[\text{DR}] = \mathbf{I} \cos \theta - [\mathbf{a} \times] \sin \theta + \mathbf{a} \mathbf{a}^T (1 - \cos \theta) \quad (20)$$

where

$$\begin{aligned} [\text{DR}] &: \text{dead-reckoning rotation matrix,} \\ \mathbf{I} &: \text{unity matrix,} \\ \theta &= |\omega| \Delta, \text{ angle of rotation in dead reckoning,} \\ \mathbf{a} &: \text{unit vector of rotation axis,} \\ [\mathbf{a} \times] &= \begin{bmatrix} 0 & -a_z & a_y \\ a_z & 0 & -a_x \\ -a_y & a_x & 0 \end{bmatrix} \\ \mathbf{a} \mathbf{a}^T &= \begin{bmatrix} a_x a_x & a_x a_y & a_x a_z \\ a_y a_x & a_y a_y & a_y a_z \\ a_z a_x & a_z a_y & a_z a_z \end{bmatrix} \end{aligned}$$

Since many aircraft simulations model coordinated turns (i.e., without sideslip), this technique may result in lower PDU rates. However, while this equation may be more accurate, this method in general needs much more computer time due to the additional computational expense of matrix multiplications.

## 7. DELAY COMPENSATION

If the simulators participating the exercise are distributed geographically, there may be significant communication delay in the received information packages. The received position/orientation need to be extrapolated to the current position/orientation during one time frame and thereafter updated at regular frame rate. Because of this "speed up" process, the updated appearance may have a severe jump in the displayed image. Hence the smoothing becomes more important as communication delay increases.

## 8. DEAD RECKONING PERFORMANCE ASSESSMENT

A complete dead reckoning procedure includes three algorithms:

- (1) Dead reckoning equation,
- (2) Smoothing equation,
- (3) Delay compensation equation,

and two parameters:

- (1) Threshold,
- (2) Smoothing points.

Variables in the DIS environment include transmission delay time and the complexity of the exercise. The delay time is a variable for different PDUs received in practice. However, in this report the delay time is assumed to be constant for all PDUs within the time period to study the effect of delay on dead reckoning. The complexity of the DIS exercise can be represented by the number of participating entities, engaged entities, and the complexity of the trajectories. The vehicles participating in the DIS exercise are divided into two groups: engaged and non-engaged entities. The engaged entities are the entities within the "area of interests" of the simulator, on which the simulator will perform dead reckoning.

The goal of this section is to demonstrate how to find the best algorithm for a given pre-described DIS environment.

### 8.1 Performance indices

The performance of the dead reckoning algorithm should be measured by considering four factors:

1. Network load,
2. Computation load of the simulator in performing dead reckoning on other entities,
3. Accuracy of the dead reckoning trajectory, compared with the true trajectory,
4. Smoothness of the dead reckoning trajectory in the visual display.

Performance indices can be defined based on each and/or combinations of the above factors. There is no single way to define these indices since they are dependent on the requirements and constraints of the individual DIS exercises. The author has defined a set of performance indices used in the software tool to be introduced in the next section. Interested readers can use them as examples.

## 8.2 The software tool

A software tool, Dead Reckoning Algorithms Assessment Tool (DRAAT), has been developed to assess the performance of the dead reckoning algorithms.<sup>8</sup> Figure 5 shows the block diagram of DRAAT. The inputs of the software are two pre-recorded trajectories of the vehicles, and an artificial trajectory files. One trajectory is the major test trajectory representing the target vehicles, the second one is the trajectory represents the pilot eye-point vehicle, and the artificial trajectory is a “race track” trajectory representing the vehicles in the background. A configuration file is read in to set the constants of the algorithms, and users can choose different algorithms interactively from the menu.

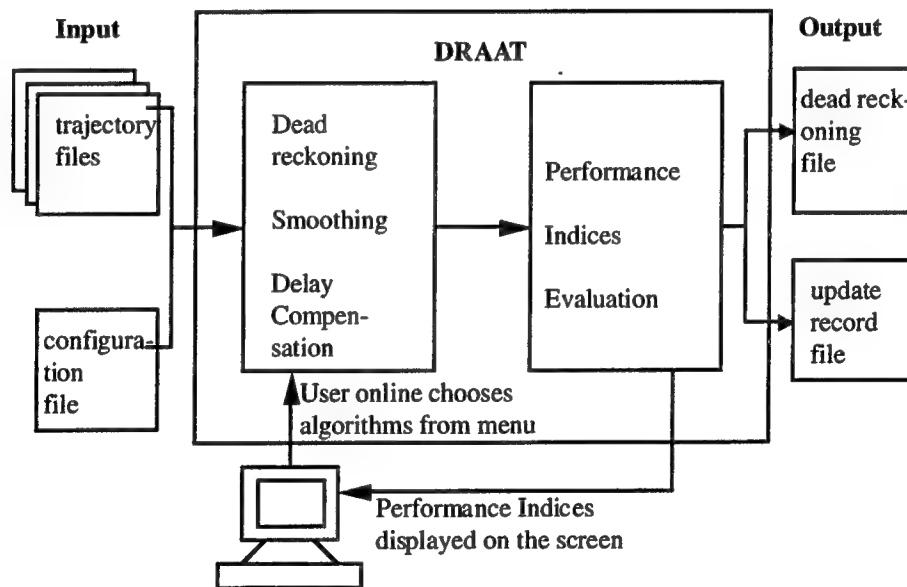


Figure 5. Block diagram of DRAAT.

The software performs dead reckoning, smoothing, and delay compensation on the input trajectories, and it records the number of PDUs generated. With the results of the evaluation, the performance indices are calculated. The outputs of the software are the performance indices, the dead reckoning trajectory of the target trajectory, and an update record file.

To calculate the performance indices, the trajectories of the engaged entities are assumed to be similar to the target trajectory in complexity. Hence, the target trajectory is used to represent all engaged entities. The calculation of computation load is based on the number of engaged entities. On the other hand, both engaged and non-engaged entities (represented by the race track trajectory) are sending PDUs to the network. Hence, the calculation of network load is based on the total number of entities. The total number of entities is a constant set in the configuration file, but the number of engaged entities is selected on-line by the user to represent the complexity of the maneuver.

The formulas for performance indices are defined as the follows:

$$\text{Network Load index (NI):} \quad \frac{\text{Actual Network Load}}{\text{Network Capability}} \quad (21)$$

$$\text{Computation Load index (CI):} \quad \frac{\text{DR Computation Load}}{\text{Computer Capability}} \quad (22)$$

$$\text{Accuracy index (AI):} \quad 1 - \frac{\text{DR Traj. RMS Error}}{\text{Maximum Error}} \quad (23)$$

$$\text{Smoothness index (SI):} \quad 1 - \frac{\text{DR total Jumpiness}}{\text{Max. Jumpiness}} \quad (24)$$

$$\text{Overall Performance Index:} \quad \frac{W1*AI + W2*SI}{W3*NI + W4*CI} \quad (25)$$

Here SI is the *S* rating described in Eq. (19), and  $W1$ ,  $W2$ ,  $W3$ , and  $W4$  are weighting constants. These weighting constants and the constants in the denominators of Eqs. (21) - (24) (i.e., network capability, computer capability, maximum error, and maximum jumpiness) are set in the configuration file.

From these formulas, we can see that the performance indices are defined to produce a better dead reckoning algorithm with smaller network load and computation load, but with larger accuracy and

smoothness indices, leading to a higher overall performance index. There is no maximum value for the overall performance index. These indices are used to make comparisons among algorithms. Their absolute values are less significant.

## 9. SUMMARY

Dead reckoning is a fundamental element of Distributed Interactive Simulation. The purpose of dead reckoning is to reduce updates required by each simulator on the network to better utilize the available bandwidth. In general, a higher order extrapolation formula provide better accuracy and lower update rate. However, extrapolation process is inherently sensitive to noise. Using an extrapolation formula of higher than second order may cause higher update rate in some cases.

Current military standard provides a list of extrapolation formulas in position/orientation. They can be found in the Annex B of the DIS standard.<sup>1</sup> Different applications have different simulation requirements; hence the needs of dead reckoning model are also different.

Smoothing is a technique to reduce the jitterness created by dead reckoning. When there is a network delay need compensation, smoothing is even more important. The tradeoff is the loss of accuracy. This report suggested a formula to evaluate the trajectory smoothness using the position jumpiness. More work need to be done in this area.

The Dead Reckoning Algorithms Assessment Tool (DRAAT) is a flexible and powerful tool to help users select a good combination of dead reckoning algorithms. More numerical experiments with different types of trajectories are still needed.

If there is an attempt in the future to reduce the Entity State PDU size, the multi-step extrapolation equations should be seriously considered.

The Summer Research Faculty has examined the existing dead-reckoning algorithms in the ITB Simulator at Wright Laboratory, and found that the whole family of the body-axes dead-reckoning algorithms were missing. After implementing the body-axes dead-reckoning algorithms, the goal of the Summer Research Faculty is achieved.

## 7. REFERENCES

1. IST, "Standard for Distributed Interactive Simulation--Application Protocols, Version 2.0," IST Report No IST-CR-94-50, March, 1994.
2. IST, "Draft, Standard for Distributed Interactive Simulation--Communication Architecture and Security," IST Report No IST-CR-94-15, March, 1994.
3. IST, "Communication Architecture for Distributed Interactive Simulation (CADIS)," IST Report No IST-CR-93-21, June, 1993.
4. Pope, A. R., "The SIMNET Network and Protocols", Report No. 7627, BBN Systems and Technologies, Cambridge, MA, June, 1991.
5. K. C. Lin and D. Schab, "Study on the Network Load in Distributed Interactive Simulation," *Proc. of the 1994 AIAA Flight Simulation Technologies Conference*, pp. 202-209, Scottsdale, AZ, August 1994.
6. K. C. Lin, M. Wang, and J. Wang, "The Smoothing of Dead Reckoning Image in Distributed Interactive Simulation," *Proc. of the 1995 AIAA Flight Simulation Technologies Conference*, Baltimore, MD, August 1995.
7. J. Burchfiel, "The Advantages of Using Quaternions Instead of Euler Angles for Representing Orientation", White Paper ASD 91-001, Third Workshop on Standards for the Interoperability of Defense Simulations, Orlando, FL, August, 1990.
8. K. C. Lin and D. Schab, "The Performance Assessment of the Dead Reckoning Algorithms in DIS," *Simulation*, Vol. 63 No. 5, pp. 318-325, November 1994.



**DYNAMIC TESTING OF F-16 BIAS AND RADIAL TIRES**

**Paul P. Lin  
Associate Professor  
Mechanical Engineering Department  
Cleveland State University  
Cleveland, OH 44115**

**Final Report for:  
Summer Research Program  
Wright Laboratory**

**Sponsored by  
Air Force Office of Scientific Research  
Bolling Air Force Base, Washington, D.C.**

**and  
Wright Laboratory**

**September, 1995**

## **DYNAMIC TESTING OF F-16 BIAS AND RADIAL TIRES**

**Paul P. Lin**  
**Associate Professor**  
**Mechanical Engineering Department**  
**Cleveland State University**

### **Abstract**

The main objective of this research was to perform three dimensional tire deformation measurement subjected to different loads, percentages of deflection and yaw angles. The optical technique used is called fringe projection. Unlike Moire Fringes, the proposed technique uses a single light source and one grating, thus requires no image superposition. As a result, the measurement is not as sensitive to vibration as the Moire method does. The other objective was to compare the magnitudes of three dimensional deformation between two types of F-16 aircraft tires made of distinct tire cords: Bias and Radial. The comparison based on some analysis is made in this report, which indicates some difference in tire deformation between these two tires.

## DYNAMIC TESTING OF F-16 BIAS AND RADIAL TIRES

Paul P. Lin

### Introduction

In non-contact measurement, several optical techniques are available. Laser ranging can yield a dense set of depth values with which surface structure can be obtained through surface fitting or approximation (Vemuri and Aggarwal, 1984). This technique, however, is usually slow and expensive. Stereo vision utilizes the disparity between the projected positions of a point in two images to infer the depth of this point (Marr and Poggio, 1976). But the correspondence between points in the stereo images is difficult to establish, and the computation is sensitive to errors introduced in digitization and camera calibration. The well known Projection Moire technique uses a white light or laser light source and two gratings of the same pitch (one in front of light and the other one in front of camera) to generate Moire interference patterns. The image is recorded in a single CCD camera, in lieu of two cameras used in stereo vision. Another very similar technique, Shadow Moire, uses only one grating near the object to generate the Moire patterns. The Moire contours thus obtained, however, do not make a difference between peaks and pits unless prior information or additional algorithms are applied. Furthermore, this technique is very sensitive to vibration due to the necessity of superimposing two images. The most accurate optical technique available today is phase-shifting interferometry (PSI). It takes time to generate three or four consecutive phase shifts to form interferograms, and thereby rendering PSI not useful for measurement of dynamic motion. This technique, by nature is also very sensitive to vibration. Another consideration is that Moire technique requires the use of two very fine pitch gratings (usually over 250 lines per inch), which makes it very difficult to visualize the generated Moire pattern on a low reflectivity aircraft tire.

The proposed fringe projection technique (Lin and Parvin, 1990; Lin, et. al., 1991) uses a single light source and a grating of 100 lines per inch in front

of light projector to generate optical fringes. No image superposition is required. In comparison with the Moire or phase-shifting technique, the fringe projection technique is less sensitive to vibration and much more computationally efficient. In 1993 Lin used an optical technique to measure the tire deformations of different aircraft bias tires: F-16, F-111 and KC-135 (Lin, et. al., '94). In 1994 the research was extended to 2-D and 3-D deformations considering the effect of yaw angle (Lin, et. al., '95). In this year (1995), dynamic testing was included for the first time. The difference in deformation between bias and radial tires is compared. Finally, the conclusions of this research and the recommendations for the future work are made.

#### **METHODOLOGY**

The measuring system consists of

- (1) Optical equipment: White light projector, grating, optical rails, CCD camera and close-range fiber-optic displacement sensor.
- (2) Image acquisition equipment: Frame grabber, image acquisition and processing software.
- (3) Data acquisition equipment: Dual-channel digital data storage oscilloscope.
- (4) High Speed Flash: Microseconds flash duration with a few nanoseconds response time.
- (5) Synchronization Device: Synchronize the tire rotation with the flash.
- (6) Recording equipment: Super VHS video recorder (VCR).
- (7) Computing equipment: 486-based micro-computer and RGB monitor.

The light produced by the white light projector passes through the grating (Ronchi ruling) and illuminate the tire surface (see Figure 1). The image captured by the CCD camera is recorded in the VCR, and then transmitted to the frame grabber where the image data are digitized and processed. The digital image is then displayed in a high resolution RGB monitor. The frame grabber and

CCD camera both have the same resolution of 512 by 480 pixels. The highest shutter speed available in this camera is 100 micro seconds. The camera, frame grabber and VCR's frame rate is 1/30 seconds. The pitch of the projection grating used was 100 lines per inch instead of 50 as used last year, and it was placed close to the tire so as to produce about 11 optical fringes when the tire was unloaded. This improvement will give higher measuring accuracy in the third dimension (i.e. Z component). Furthermore, the use of a new CCD camera with 2/3 inch imager made it possible to be 1/3 closer to the tire when comparing with the old camera with 1/2 inch imager. This arrangement is particularly important for a low reflectivity object such as tires.

The principle of the 3-D optical measurement used here is based on the curvature change of projected fringes and the spacing between two adjacent fringes. The first step in image analysis is to accurately detect the locations of fringe centers, line by line. Usually the fringe center locations are limited to the so-called one-pixel resolution which is provided by the CCD camera. In this research, the so-called sub-pixel resolution was employed in order to improve the fringe center detection accuracy (Lin and Parvin, '90). It should be noted that three dimensional geometry determination greatly depends upon the accuracy of fringe centers. When performing image analysis, fringe centers are scanned from top to bottom and left to right (see Figure 2).

It is necessary to specify the location of the reference point within the tire and near the wheel flange. In addition, a reference plane (xy plane) passing through this point and perpendicular to the viewing direction has to be established. It is worthwhile to note that when a tire is loaded, not only the location (x and y components) of the reference point changes, the height (z component) of the point (i.e. perpendicular to the sidewall) changes as well. In this study, a close-range fiber-optic sensor was installed to accurately measure the reference point's height change when a tire is loaded. During the tire loading process, this sensor moved with the tire in order to keep the detecting position constant.

A wobbling motion usually occur when a tire rotates against a flywheel. However, it was found that the degree of wobble was very small at the tire test

facility (dynamometer 120). Thus, there was no need to synchronize the tire rotation with a flash. A projector producing continuous light and a CCD camera set at a shutter speed of 1/1000 seconds were used this year. A synchronization device was prepared just in case if wobbling ever occurs. It was designed to turn on the high speed flash exactly once per revolution so that same points of interest in the tire can be captured by the CCD camera. A reflective tape is placed on the reference point and the probe of the fiber-optic sensor is then focused on the tape. Initially, the height of this point is detectable. As the tire rotates, the tape goes away from the fiber-optic light beam which results in a out-of-range detection with zero output voltage. As the tire comes back to the same angular position (i.e. exactly one revolution), the voltage will be back to the initial highest value. Since the voltage generated by the fiber-optic sensor changes so fast that it is necessary to use a digital data storage oscilloscope to store the digital data for later examination. A threshold voltage near the peak value can be set so as to trigger the flash. Thus, the flash is fired only once per revolution. The flash takes a few nanoseconds of response time and 9 microseconds of flash duration operating at the power of 60 Watts. The synchronization device worked well. The only concern was that the light energy reduces when the flashing frequency increases. In many cases, the acquired images appeared to be too dark to be used for image analysis.

The acquired images were filtered and analyzed. Figs. 3 and 4 show the images containing optical fringes when the radial tire was subjected to 0% deflection at 0 mph (miles per hour) and 30% deflection at 40 mph, respectively. To determine the three dimensional geometry of tire deformation, the in-house developed computing algorithms were employed.

### Results and Discussion

In this research, F-16 bias and radial tires were tested under the same loading conditions statically and dynamically. Both tires were loaded against a flywheel with 30% deflection at 0°, 2°, ±4° yaw angles. The tire rotating speeds are 10 mph and 40 mph at the various yaw angles. Furthermore, the tires were tested under the conditions of maximum-power takeoff, landing-taxi and taxi-refused takeoff at 0° yaw angle. Table I shows that both tires require different loads to produce the same deflection. In the tests of maximum-power

takeoff, landing-taxi and taxi-refused takeoff, both tires followed the same load and speed profiles which were preprogrammed. In order to have a meaningful comparison between two tires, the initial pressure on the radial tire was raised to 390 psi.

TABLE I

Loading Comparison between F-16 Bias and Radial Tires

Loading Condition: Rated Pressure (310 psi initially)  
Corrected Load (as shown below)

*<Subjected to Flywheel Loading only>*

(A) Static, 10 and 40 MPH

Deflection	Bias	Radial
30 %	14000 lb (initially 310 psi)	12000 lb (initially 310 psi)

(B) Maximum-Power Takeoff

Deflection	Bias	Radial
33 %	16200 lb (initially 310 psi)	16200 lb (initially 390 psi)

It can be seen in Table I that the radial tire requires about 16% less load than the bias tire for producing 30% deflection. In other words, when subjected to the same load, it is easier to deflect a radial tire. The 3-D graphs included in the Appendix show only the 3-D reconstructed geometry of a small but significant portion of a deformed tire (a rectangular region near the contact between tire and flywheel).

The complete analysis of yaw angle effect on tire 3-D deformation is not available yet. However, the reference point's height changes subjected to different yaw angles and percentages of tire deflection were measured directly with the fiber-optic sensor. The selected reference point was located on the tire sidewall, near the wheel flange and approximately at 11 O'clock position. The height change is perpendicular to the tire sidewall in the direction parallel to the wheel axle, the so-called out-of-plane deformation. TABLE II shows the measured height change data comparison.

TABLE II

Comparison of the Reference Point's Height Changes

Unit: mm

Note: Negative values correspond to inward deformation at the reference point located at 11 O'clock position.

(A) Yaw angle = 0°	Radial	Bias
30% deflection, static	-0.38	-0.63
30% deflection, 10 mph	-0.72	-0.87
30% deflection, 40 mph	-2.12	-1.99
<hr/>		
(B) Yaw angle = 2°	Radial	Bias
30% deflection, static	-0.71	-0.70
30% deflection, 10 mph	-1.00	-0.95
30% deflection, 40 mph	-2.50	-2.16
<hr/>		
(C) Yaw angle = 4°	Radial	Bias
30% deflection, static	-1.02	-0.77
30% deflection, 10 mph	-1.63	-1.05
30% deflection, 40 mph	-3.48	-3.12
<hr/>		
(B) Yaw angle = -4°	Radial	Bias
30% deflection, static	-0.47	-0.15
30% deflection, 10 mph	-0.67	-0.46
30% deflection, 40 mph	-1.94	-0.83
<hr/>		

Table II indicates that the reference point's height change is a little more for the radial tire than for the bias tire. For both tires, the height change gets larger as the yaw angle increases, and gets larger with positive yaw angles than with negative ones.

As can be seen in the Appendix, when subjected to the same percentage of deflection, the deformation magnitudes appear to be about the same between these two tires, but the shape of deformation is different. Unlike the bias tire, the radial tire looks like a flat tire when loaded.

The fiber-optic sensor indicates that the tire's reference point deforms more inward as the tire rotating speed increases. The centrifugal force



generated by tire rotation tends to push the air elsewhere and adds an additional load on the tire. As a result, the tire portion near the contact area deforms more outward as the tire speed increases.

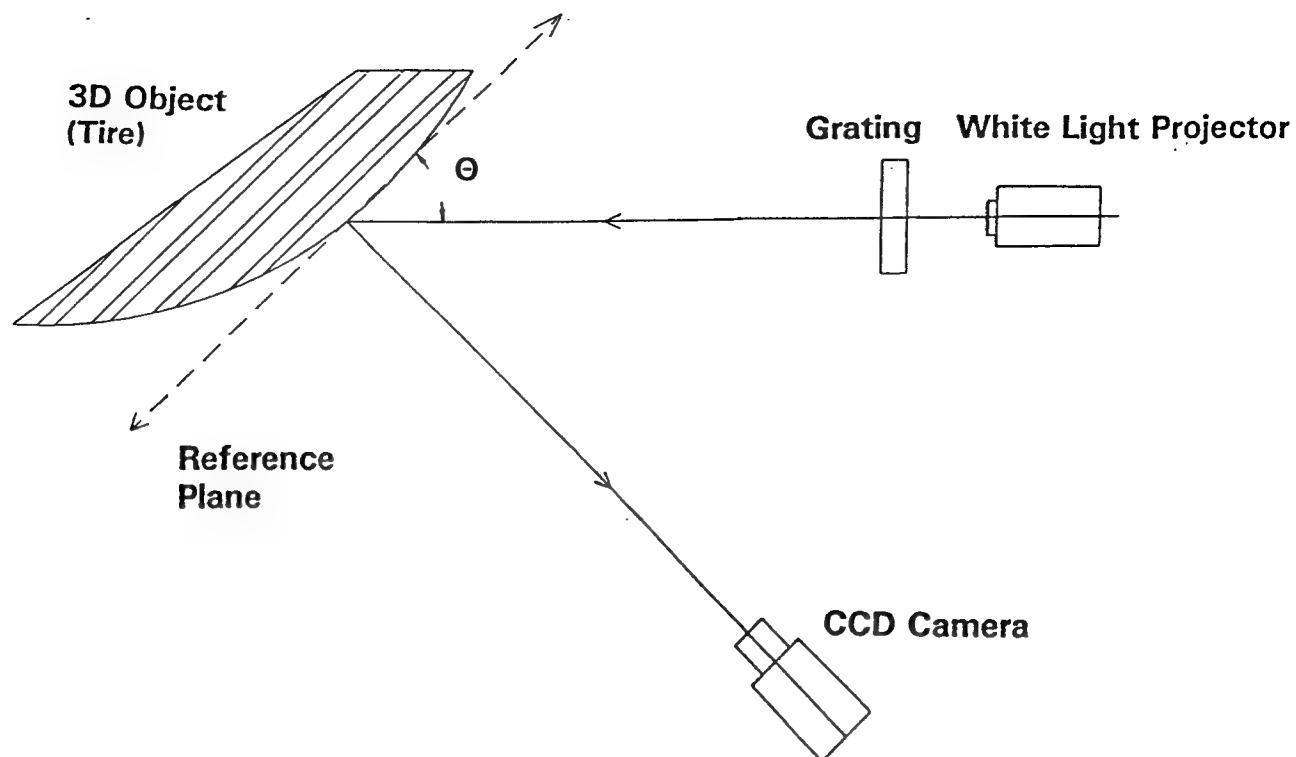
### Conclusions and Future Work

The measuring system and the optical technique used worked very well. The selected camera shutter speed of 1/1000 seconds was adequate. The results of preliminary image analysis seem to indicate that when subjected to the same percentage of deflection, the radial and bias tires exhibit approximately the same deformation magnitudes, even though the shapes of deformation are different. The follow-up project will focus on the image analysis for all tests conducted this summer, then summarize all the test data and place them in a deformation data matrix. Future work will include

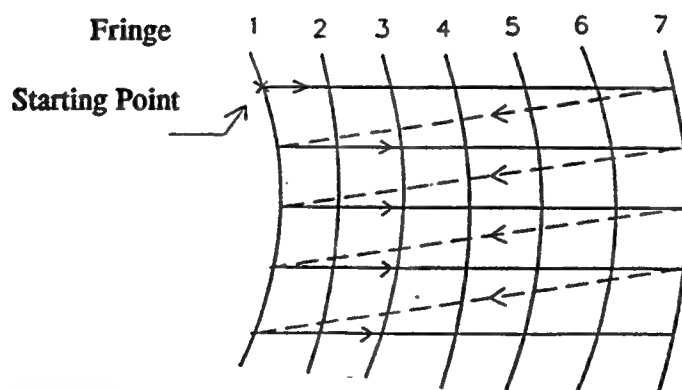
- (1) testing of tire deformation under various braking conditions.
- (2) building an artificial neural network to establish the mapping relationship between inputs and outputs.
- (3) using Taguchi method to determine the significant factors contributed to tire deformation.

### References

- Lin, P. P. and Parvin, F., (1990) "Edge Detection with Subpixel Resolution and its Application to Radius Measurement via Fringe Projection Technique," SME Technical Paper, MS90-576, pp. 4-13 - 4-27.
- Lin, P. P., Parvin, F., and Schoenig, Jr., F. C., (1991) "Optical Gaging of Very Short-term Surface Waviness," Transactions of NAMAR/SME, pp. 327-322.
- Lin, P. P., Chawla, M. D., and Ulrich, P. C., (1994) "Optical Technique for Measuring Tire Deformation and Strains - Preliminary Results," SAE Technical Paper No. 941178, Aerospace Atlantic Conference and Exposition.
- Lin, P. P., Chawla, M. D., and Wagner, P. M., (1995) "Deformation Comparison between Bias and Radial Aircraft Tires Using Optical Techniques", SAE Technical Paper No. 951433, Aerospace Atlantic Conference and Exposition.
- Marr, D. and Poggio, T., (1976) "Cooperative Computation of Stereo Disparity," Science, V. 194, pp. 283-287.
- Vemuri, B. C. and Aggarwal, J. K., (1984) "3-Dimensional Reconstruction of Objects from Range Data," Proc. of 7th Int. Conf. on Pattern Recognition, V1, pp. 752-755.



**Fig.1 Experimental Arrangement for 3D Geometry Measurement**



**Note:** Fringe 1 is scanned first (from top to bottom) and the remaining fringes are then scanned horizontally line by line as shown in this figure.

**Fig. 2 Image Data Scanning Process**

F16 RADIAL

FZ= 0.30

TP=310.50

FV= 0.00

IT= 91.00

IR= 91.25

30AG95 13:09:17

Fig. 3

F16 RADIAL

FZ= 11.95

TP=317.00

FV= 40.20

TT= 94.00

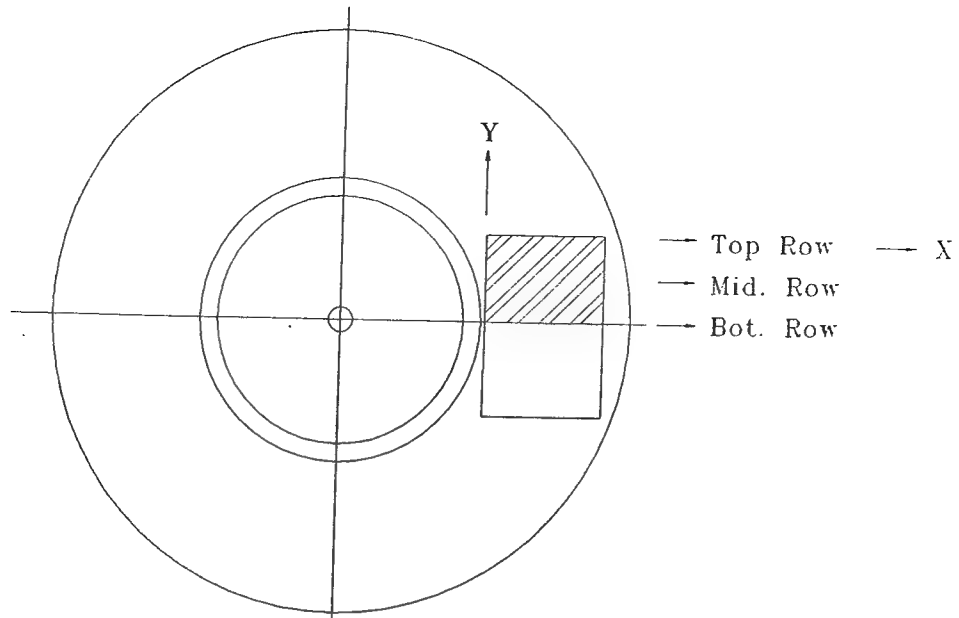
IR= 96.50

30AG95 13:36:05

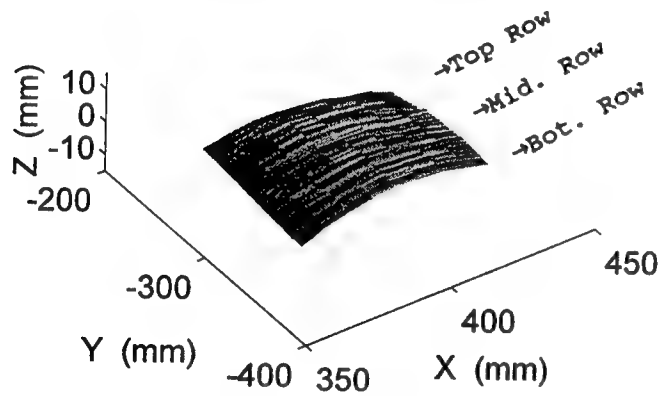
Fig. 4

## **APPENDIX**

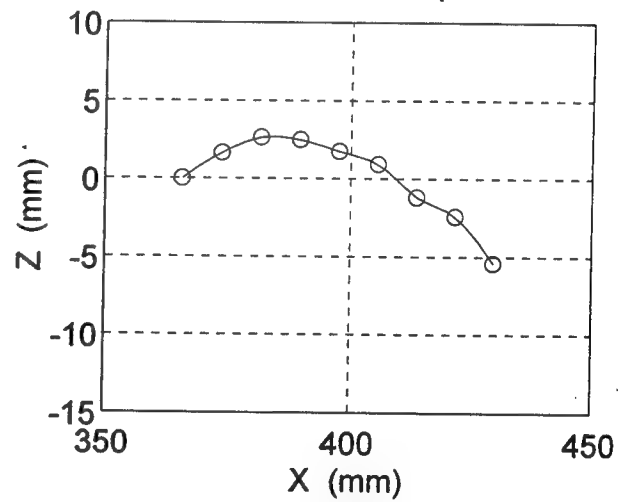
### **Selected Results of 3-D Tire Deformation**



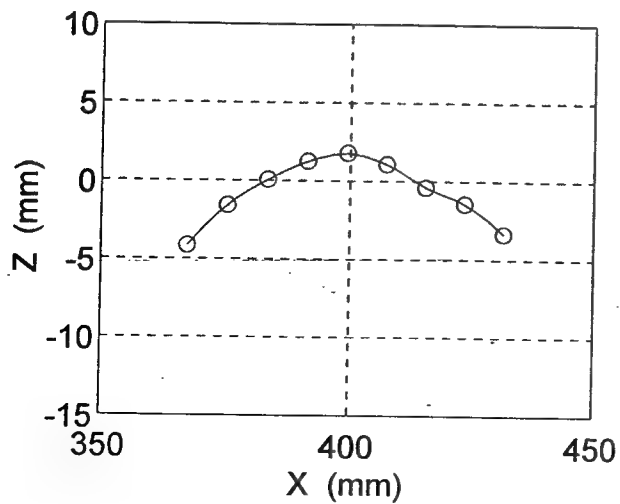
F16 Bias: 0% Deflection, 0 mph  
Initial Tire Pressure: 310 psi



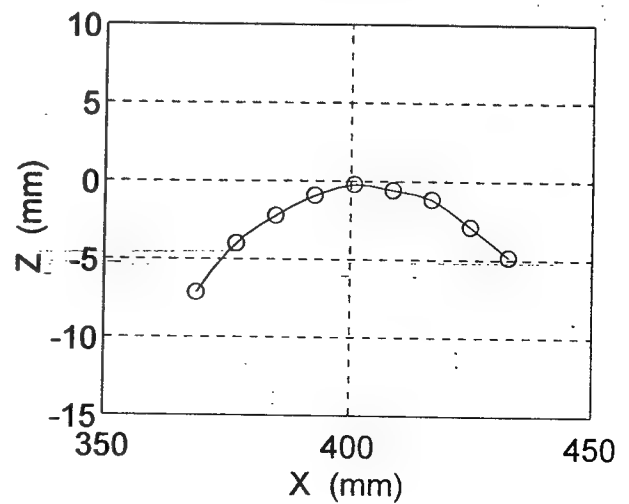
Deformation at Top Row



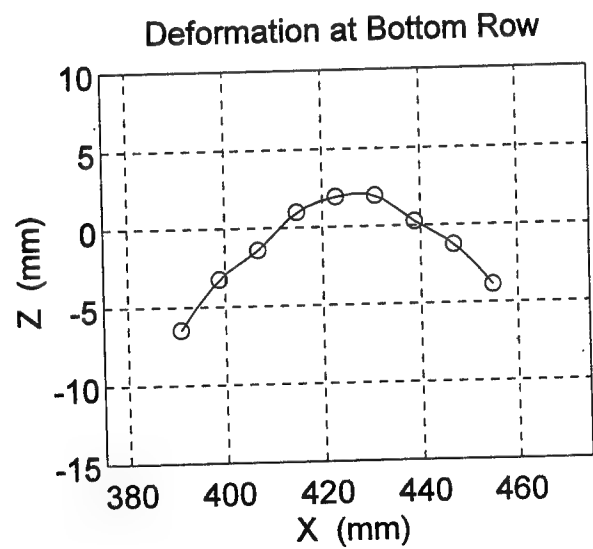
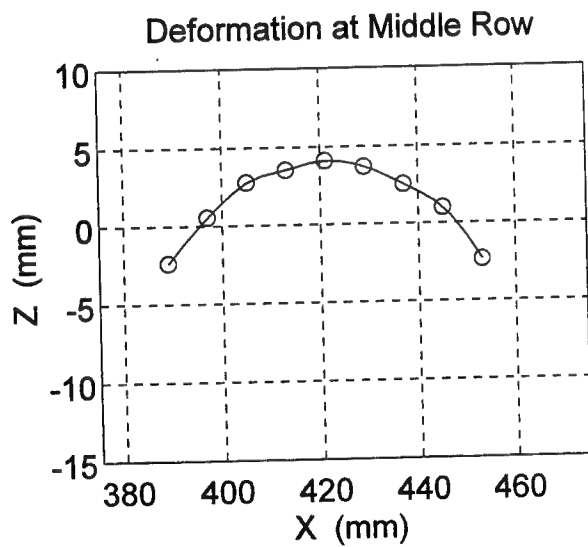
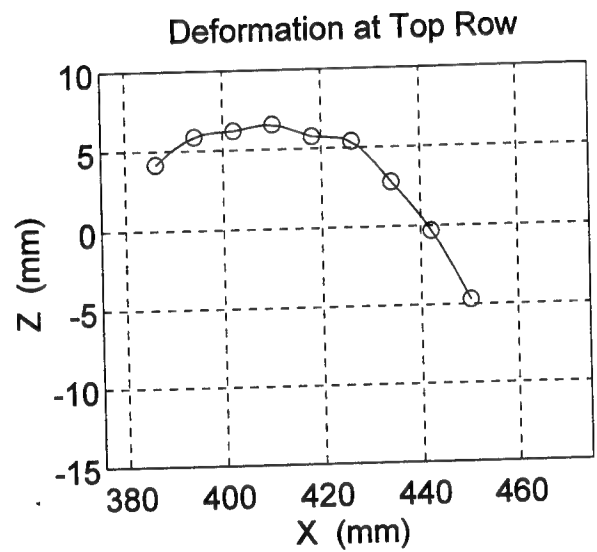
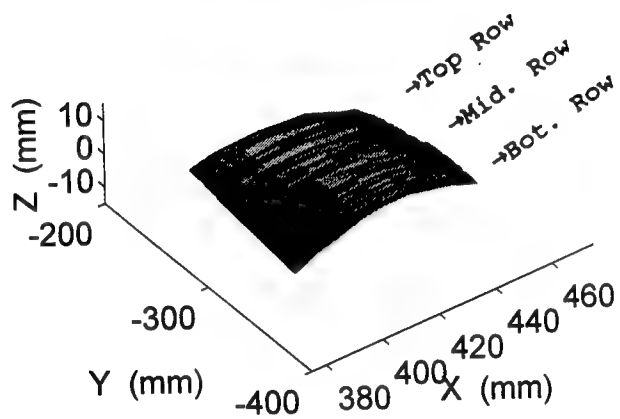
Deformation at Middle Row



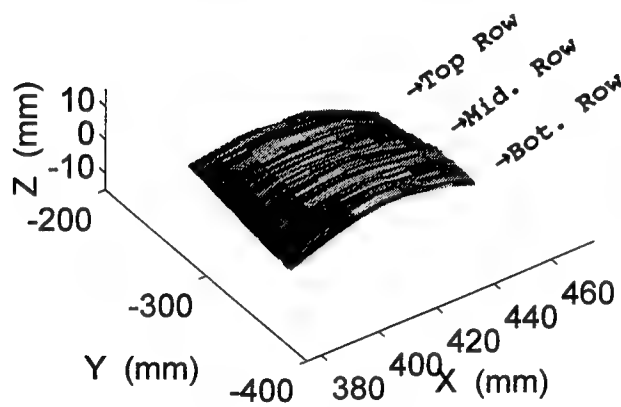
Deformation at Bottom Row



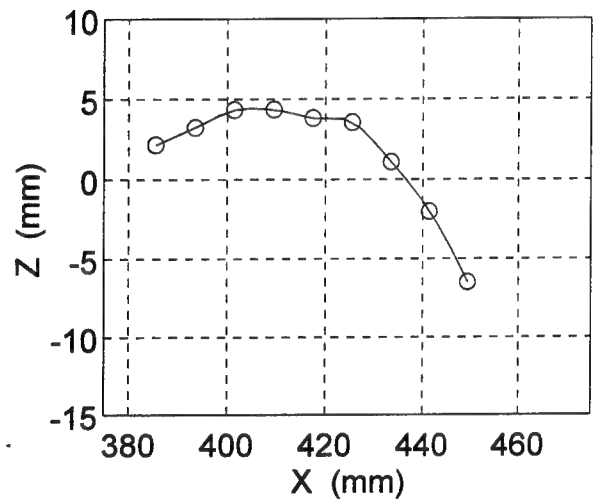
F16 Bias: 30% Defl., 0 mph  
Load: 12000 lb, Pressure: 317 psi



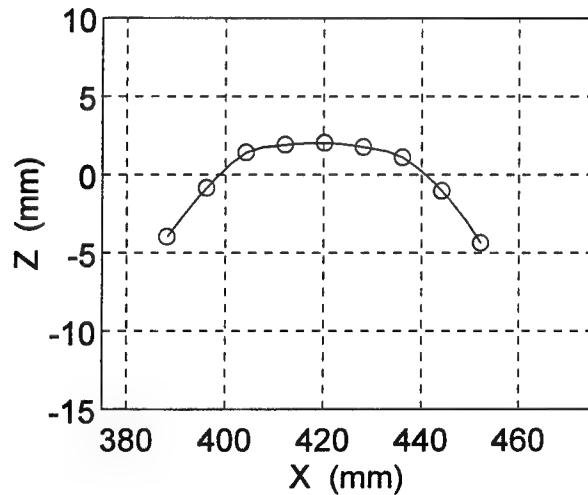
F16 Bias: 30% Defl., 40 mph  
Load: 12000 lb, Pressure: 320 psi



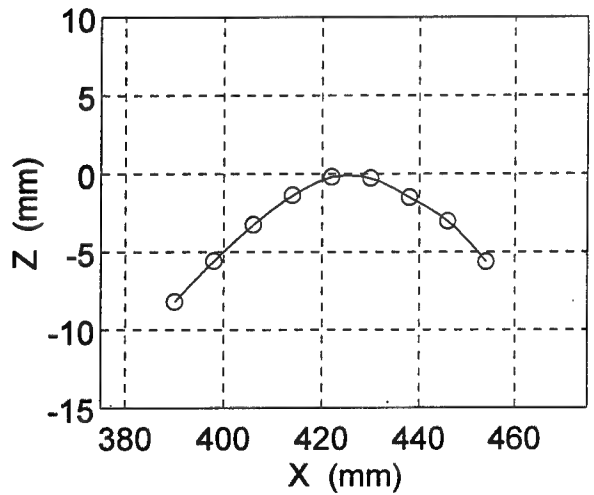
Deformation at Top Row



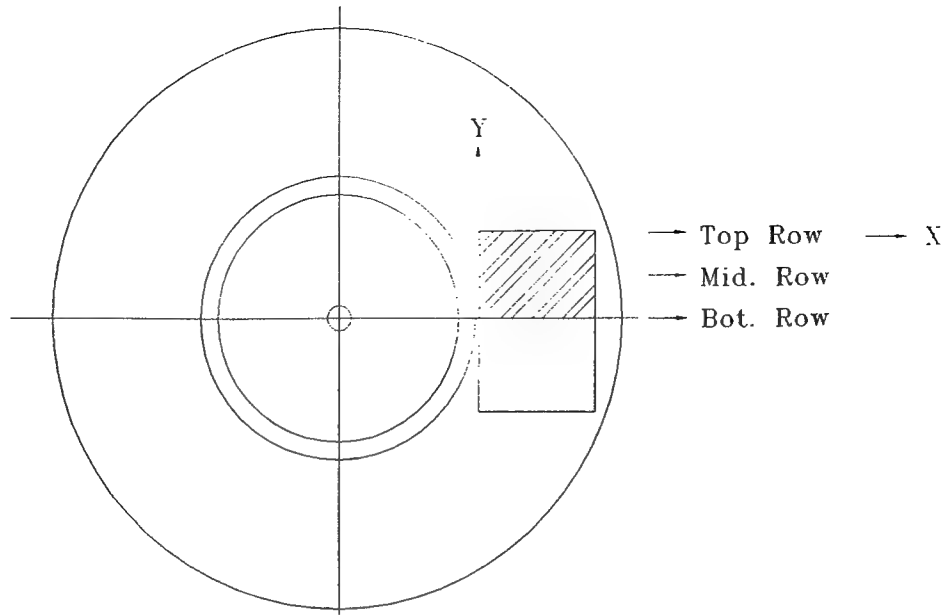
Deformation at Middle Row



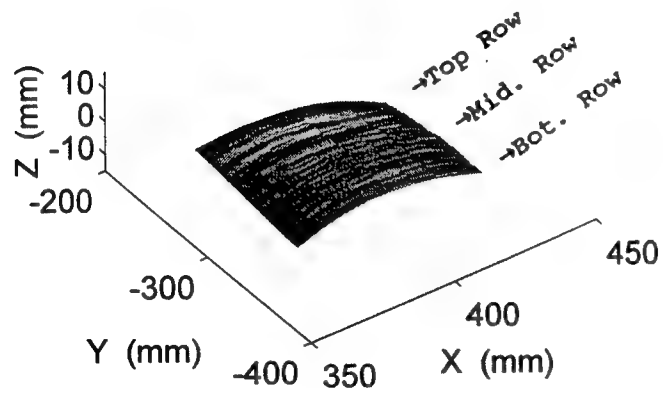
Deformation at Bottom Row



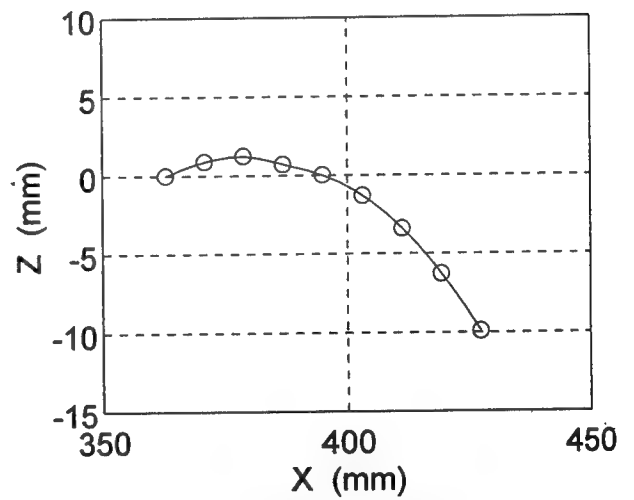




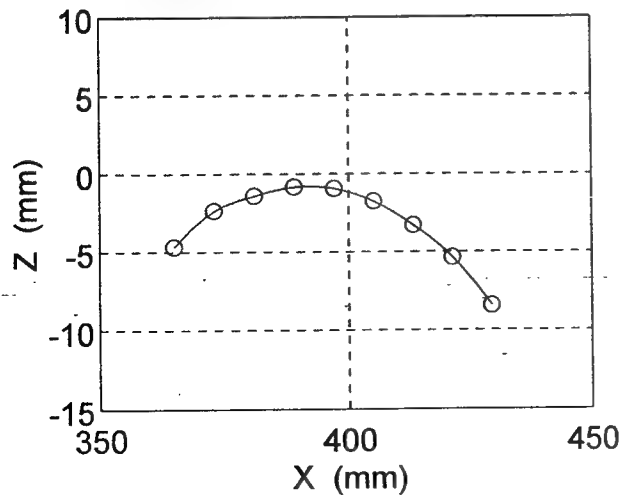
F16 Radial: 0% Deflection, 0 mph  
Initial Tire Pressure: 310 psi



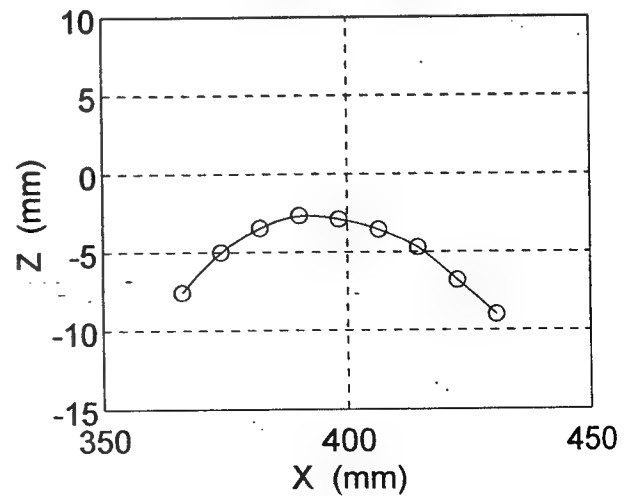
Deformation at Top Row



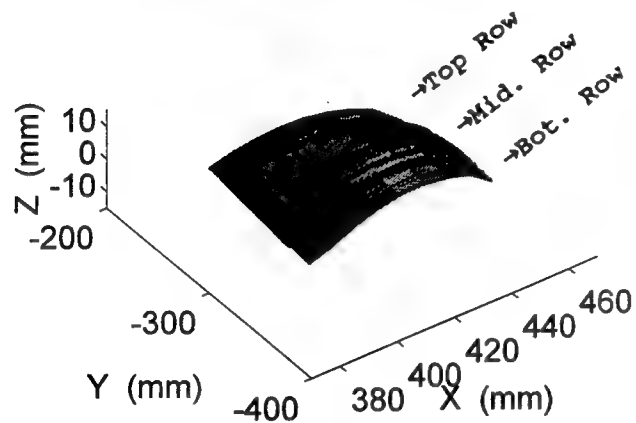
Deformation at Middle Row



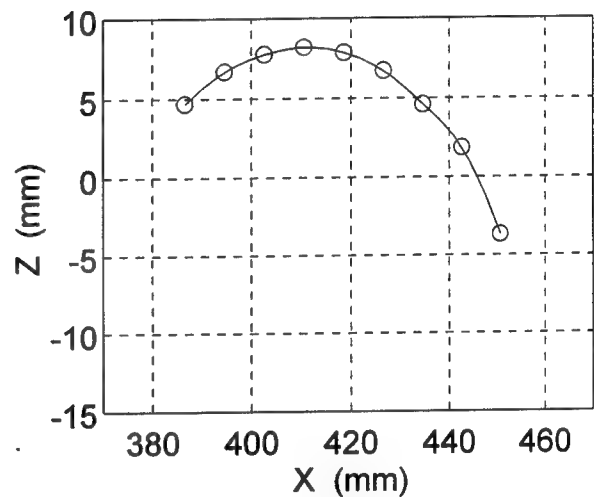
Deformation at Bottom Row



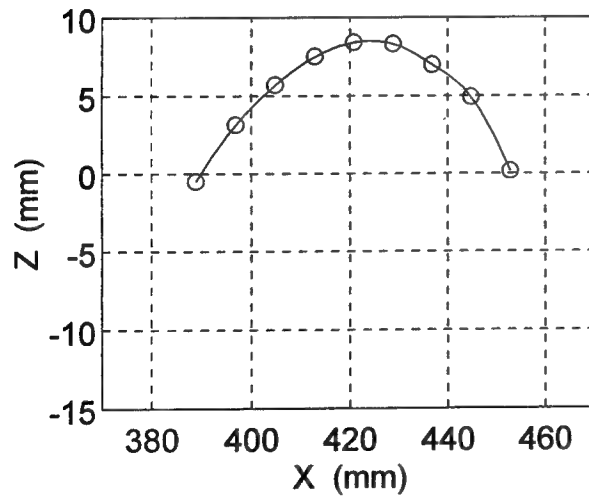
F16 Radial: 30% Defl., 0 mph  
Load: 12000 lb, Pressure: 317 psi



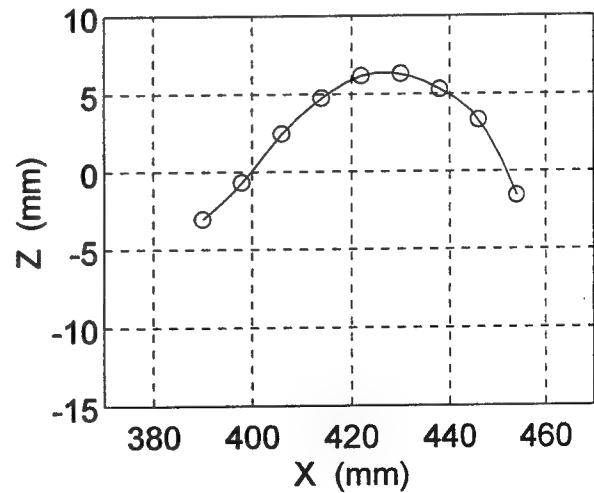
Deformation at Top Row



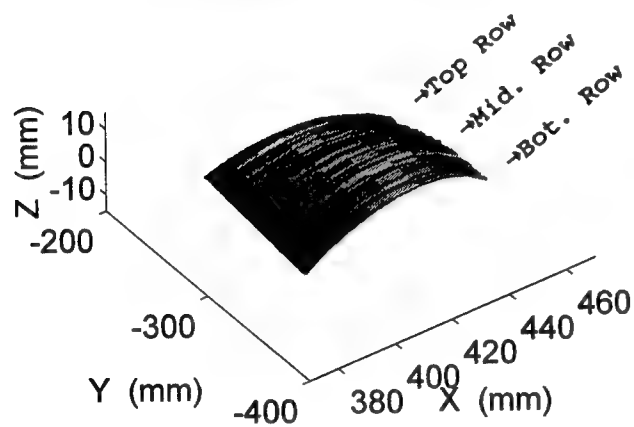
Deformation at Middle Row



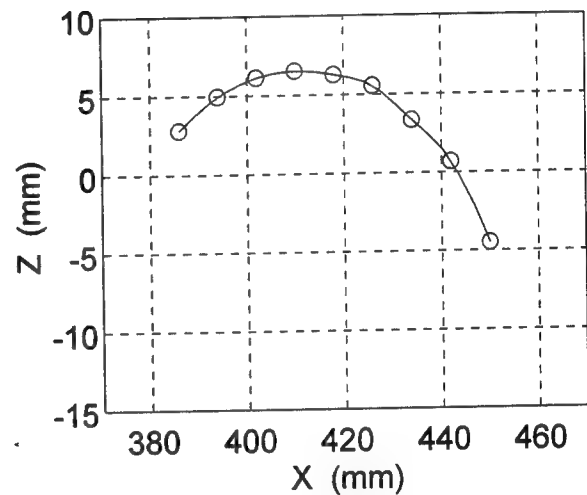
Deformation at Bottom Row



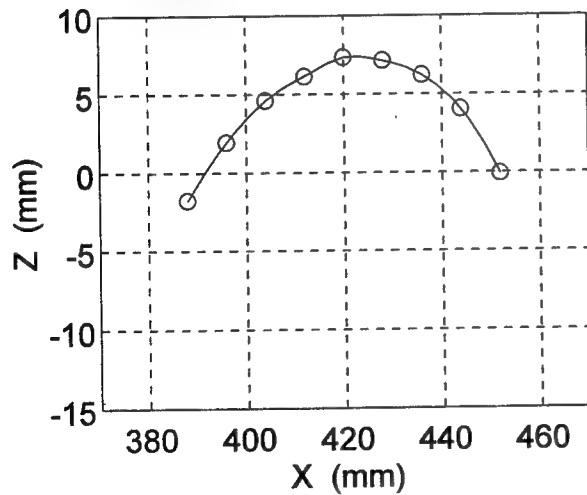
F16 Radial: 30% Defl., 40 mph  
Load: 12000 lb, Pressure: 317 psi



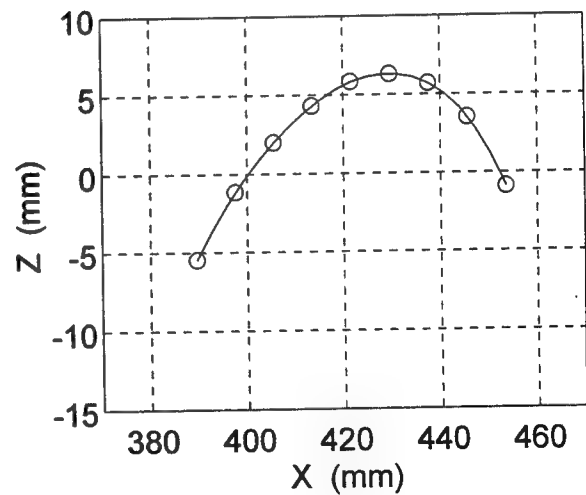
Deformation at Top Row



Deformation at Middle Row



Deformation at Bottom Row



**Studies of Tricresyl Phosphate(TCP)  
and Its Iron Mixture Using  
Differential Scanning Calorimeter(DSC)**

**Christopher C. Lu  
Associate Professor  
Depart of Chemical and Material  
Engineering**

**University of Dayton  
300 College Park Avenue  
Dayton, OH 45469**

**Final Report for:  
Summer Research Program  
Research & Development Laboratories**

**Sponsored by:  
Air Force Office of Scientific Research  
Wright Patterson Air Force Base  
Dayton, OH**

**July 1995**

**Studies of Tricresyl Phosphate(TCP)  
and Its Iron Mixture Using  
Differential Scanning Calorimeter(DSC)**

**Christopher C. Lu  
Associate Professor  
Department of Chemical and Material  
Engineering  
University of Dayton**

**Abstract**

When TCP solution(10mg ~ 20mg) was heated at the heating rate of 10°C/min under the air (or the oxygen) environment, flow rate of 50ml/min, first an exothermic oxidation took place between 200°C and 280°C, followed by an endothermic evaporation occurring at about 320°C, and followed by another exothermic char formation at about 350°C. Finally, the thermal decomposition of char would occur at about 480°C. A sample collected at 250°C had about 50% of TCP (or soluble organic compounds)left, and that collected at 325°C only had 5% TCP left.

When iron powder (about 7mg) was heated at the same conditions described above, there always showed an exothermic peak, at about 300°C under the nitrogen, and about 600°C under the air. These exothermic peaks are probably contributed by the change of crystal structures from alpha or delta to gamma.

**Studies of Tricresyl Phosphate(TCP)  
and Its Iron Mixture Using  
Differential Scanning Calorimeter(DSC)**

**Introduction**

The main purpose of this research is using Du Pont 910 Differential Scanning Calorimeter(DSC) to study the thermal behavior of tricresyl phosphate (TCP) which has been used as a hydraulic fluid and as an antiwear additive to various lubricating fluids. The differential scanning calorimetry is a technique which has been used to study the heat and temperature transitions, such as the boiling point, heat of vaporization (endothermic), and oxidation reaction (exothermic), of materials under variable heating environments.

This report mainly covers two subjects; the first subject discusses the TCP solution alone, and the second subject discusses the mixture of TCP solution with iron powder. For the first subject, the TCP solution was heated in DSC for air (or oxygen) flow rate of 50ml/min, the peaks of the DSC thermogram were used to study exothermic, and endothermic reactions, and were used to determine the boiling point as well as the heat of vaporization. The oxidized residues were also analyzed using Hewlett Packard 1050 Liquid Chromatograph (HP-1050 LC) to determine the remaining amount of TCP or soluble organic compounds during an oxidation process. For the second subject, the TCP/iron mixture was heated under the similar operating condition as described in the first task. The residues were also examined using HPLC and SEM (scanning electron microscopy).

**Results and Discussions**

**[I] TCP Solution**

Durad 125 TCP solution was placed in an aluminum pan as a sample for the DSC experiment, and three(3) pan sample configurations were

used in this study: (a) air flow for an open pan, (b) oxygen flow for an open pan, and (c) oxygen flow for a sealed one pinhole pan.

#### (1) Air Flow For Open Pan

TCP solution was heated from the room temperature to 550°C at a rate of 10°C/min under the air flow rate of 50ml/min. Figure 1 shows that the endothermic evaporation takes place at about 320°C, and the area of the peak (heat of vaporization) is integrated by a DSC software as a value of about 138 joule/gm (33 cal/gm); an exothermic reaction follows at about 350°C for about 31 joule/gm of heat evolution.

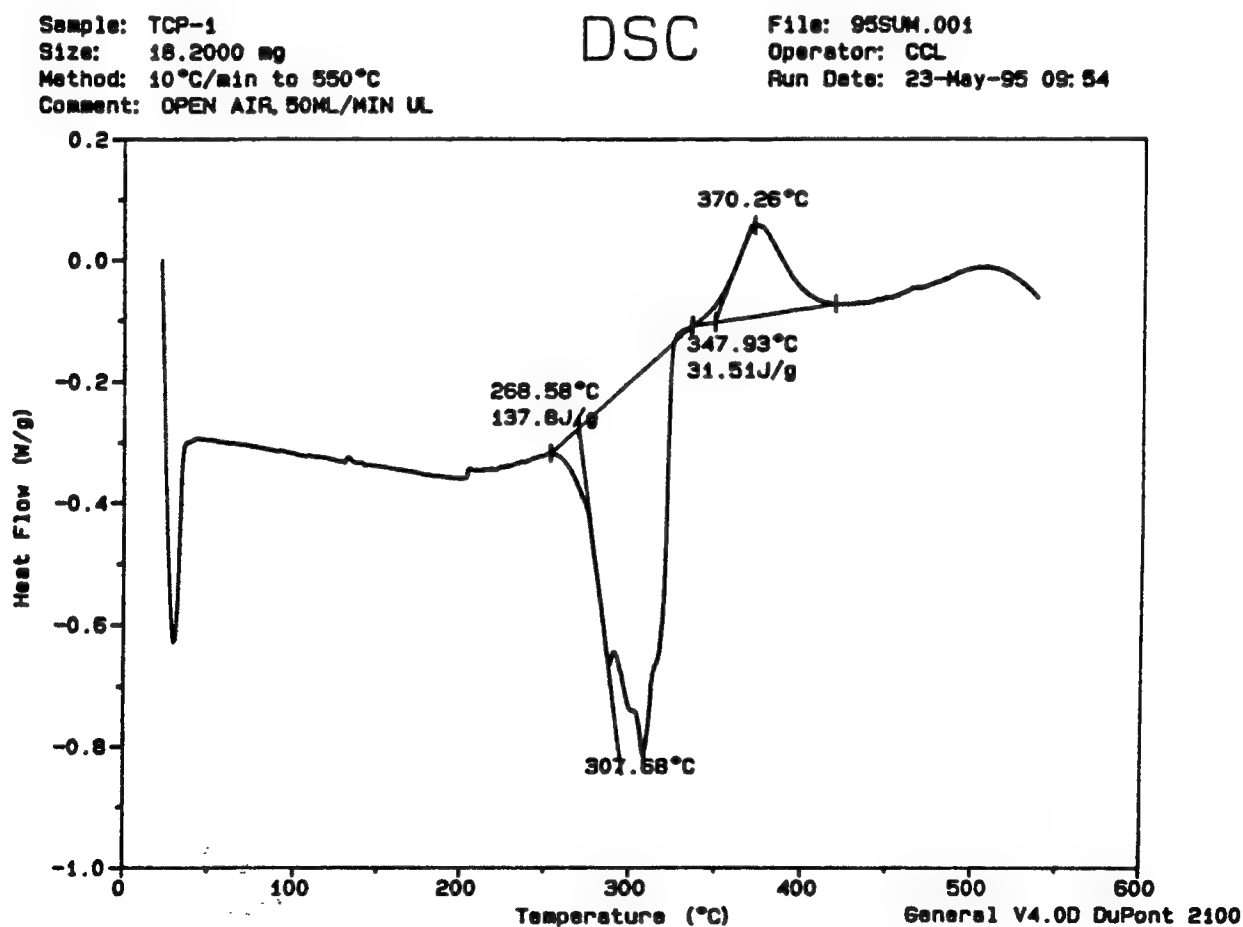


Fig.1 Air Flow For Open Pan

The boiling point of TCP was estimated as 265°C at 10 mm Hg, and

the heat of vaporization was 44.5 cal/g (Ref. 1); the boiling point of p-tricresyl phosphate was given as 410°C at the atmospheric pressure (Ref.2). (Please note that TCP is a mixture of isomers). Studies of oxidation and molecular weight distribution for TCP solution were conducted using Thermogravimetric Analysis (TGA), sealed tubes, and air bubble methods (Ref. 3,4). The study discussed that when a TCP solution was heated, a combination of oxidation, evaporation, and polymerization occurred between 180°C and 400°C. At 180°C to 300°C, the oxidation gradually increased the molecular weight, and highly increased between 330°C and 380°C; and finally formed darkish char at about 450°C by thermal decomposition. In the figure 1, a small exothermic peak (between 200°C and 250°C) suggests the beginning of oxidation reaction, followed by an endothermic evaporation, and then by another exothermic peak of a combination of oxidation and polymerization reactions.

## (2) Oxygen Flow For Open Pan and Sealed-One Hole Pan

TCP solution was heated under oxygen flow rate of 50ml/min for both open pan and sealed pan with one 0.65 mm pinhole on the lid. The DSC thermogram are depicted in figures 2,3. From these figures, it is more clear that the exothermic oxidation peak occurred between 200°C and 280°C prior to the evaporation process. The heat of vaporization under the pure oxygen environment is 73 joule/gm comparing with 138 joule/gm under the same air flow rate. It is probably that more TCP has been oxidized under the oxygen environment prior to evaporation. The heat evolution followed the evaporation is doubled under oxygen against under the air (63 j/g, and 32 j/g), because of the higher oxygen content promotes both oxidation and polymerization reactions. When the TCP pan is sealed with an one pinhole lid (figure 3), the amount of oxygen introduced into the system and the amount of vapor escaping from the system is restricted, and these factors will affect the processes of oxidation, evaporation, and polymerization. Both of heat of vaporization and the followed heat of oxidation/polymerization are



larger for the sealed pan with a pinhole condition than for the open pan case.

Sample: TCP/02 (50CC/MIN), Kinetic Study  
Size: 21.8000 mg  
Method: 10°C/min to 550°C  
Comment: OPEN 02, 10/MIN TO 550

DSC

File: 95SUM.031  
Operator: LU  
Run Date: 13-Jun-95 09:11

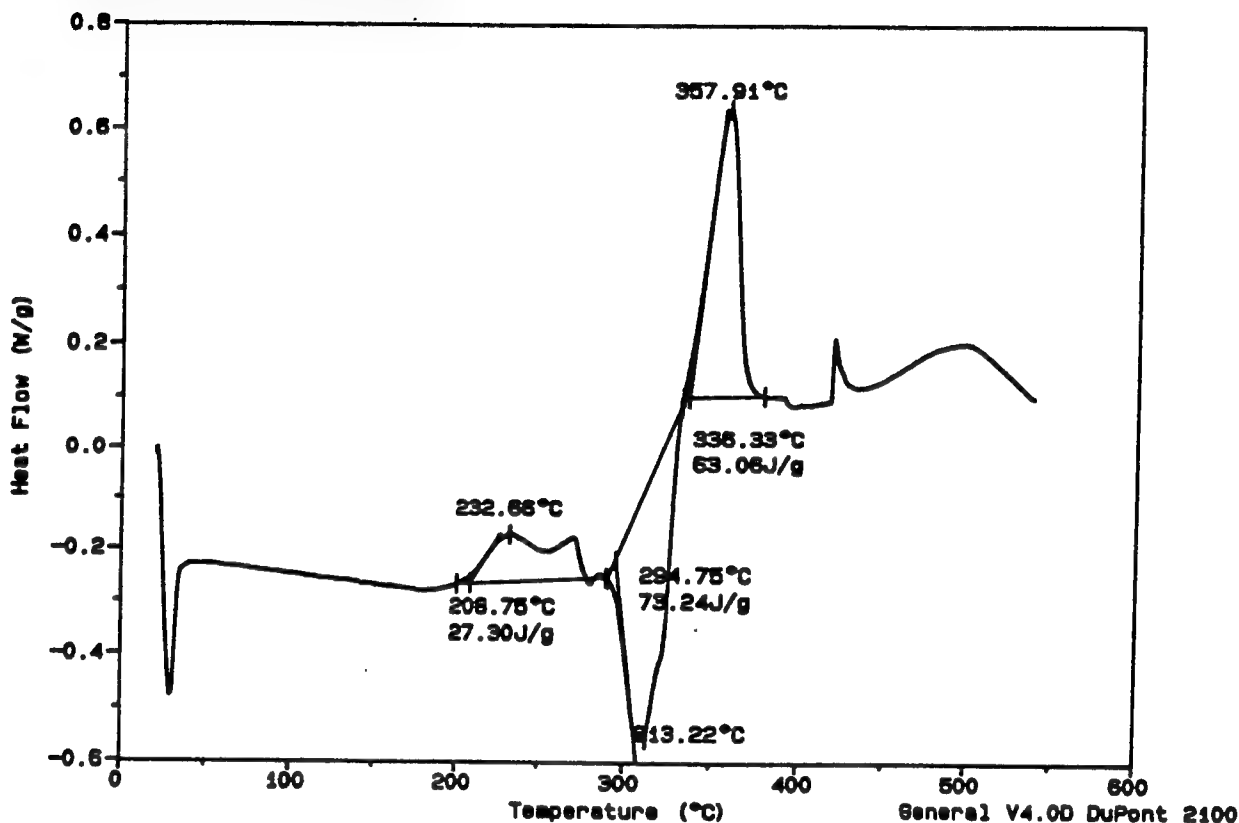


Fig.2 Oxygen Flow for Open Pan

It is possible that the amount of oxygen admitting to the system through a small pinhole is already enough to smoothly carry on the oxidation and polymerization reactions at this heating rate, but the escape of gases is limited by the small hole and becomes an important factor. The reaction mechanism of lubricants' oxidation and the diffusion of oxygen/vapors through a pinhole is discussed by Zhang et.al.(Ref.4). In the figure 3, it also shows a large amount of exothermic reaction taking place at about 480°C for the thermal decomposition of char residues.

Sample: TCP/02, SEALED, 1HOLE, KINETIC  
Size: 16.8000 mg  
Method: 10°C/min to 550°C  
Comment: SEAL PAN, 1 HOLE, O2 50ML/MIN, 10/MIN TO 550

DSC

File: 95SUM.033  
Operator: LU  
Run Date: 13-Jun-95 14:15

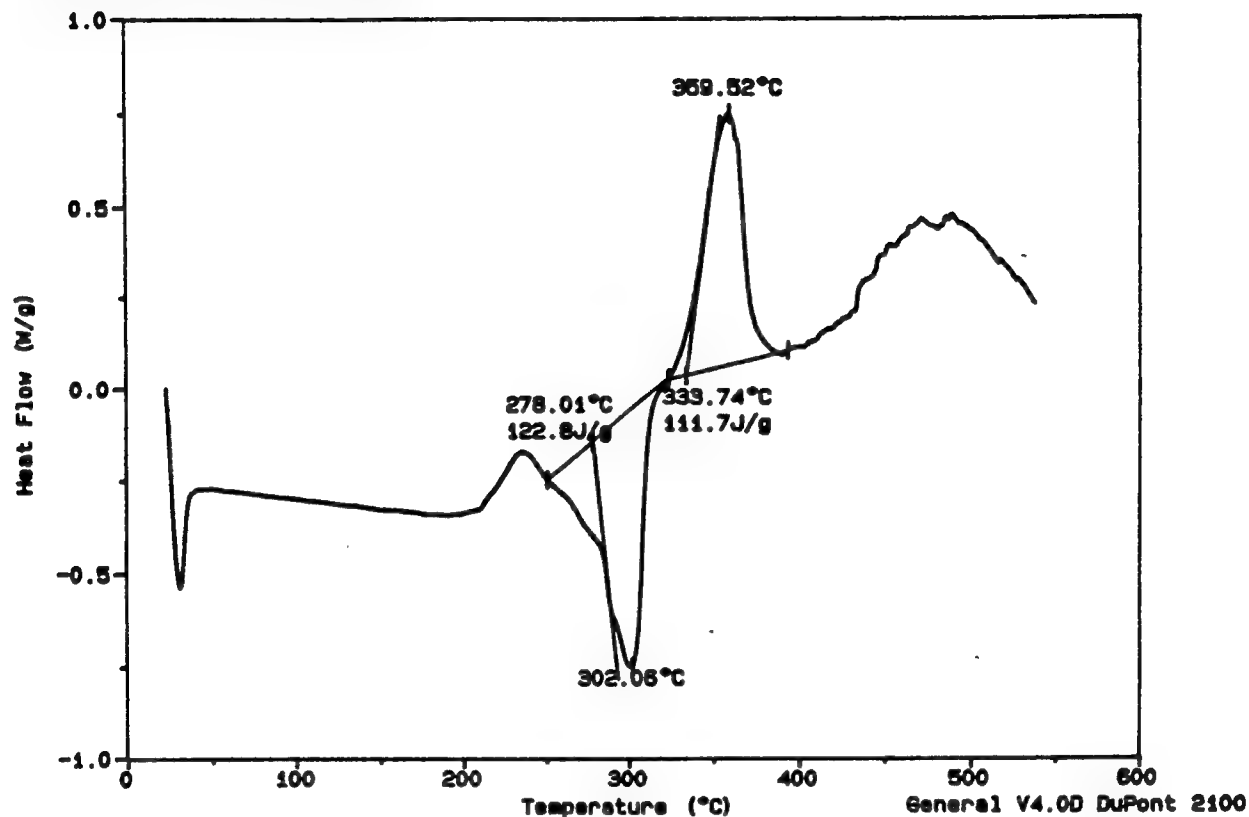


Fig.3 Oxygen Flow For Sealed-One Hole Pan

### (3) Quantitative Analysis

Hewlett Packard 1050 Liquid Chromatograph (HP 1050) was used to analyze the residues of the oxidized TCP solution, and to measure the non-reacted TCP left in the product. To do this, an amount of 10 $\mu$ l TCP solution was put in an one pinhole sealed pan, and the sample was heated to 250°C, 325°C, 380°C, and 400°C separately according to the trend of a DSC thermogram shown in figure 4.

Sample: TCP/02, SEALED, 1HOLE, KINETIC  
Size: 11.5700 mg  
Method: 10°C/min 550°C  
Comment: SEAL PAN, 1 HOLE, 02 50ML/MIN, 10/MIN 550 CHECK RUNS 038, 041

DSC

File: 95SUM.042  
Operator: LU  
Run Date: 20-Jun-95 11:33

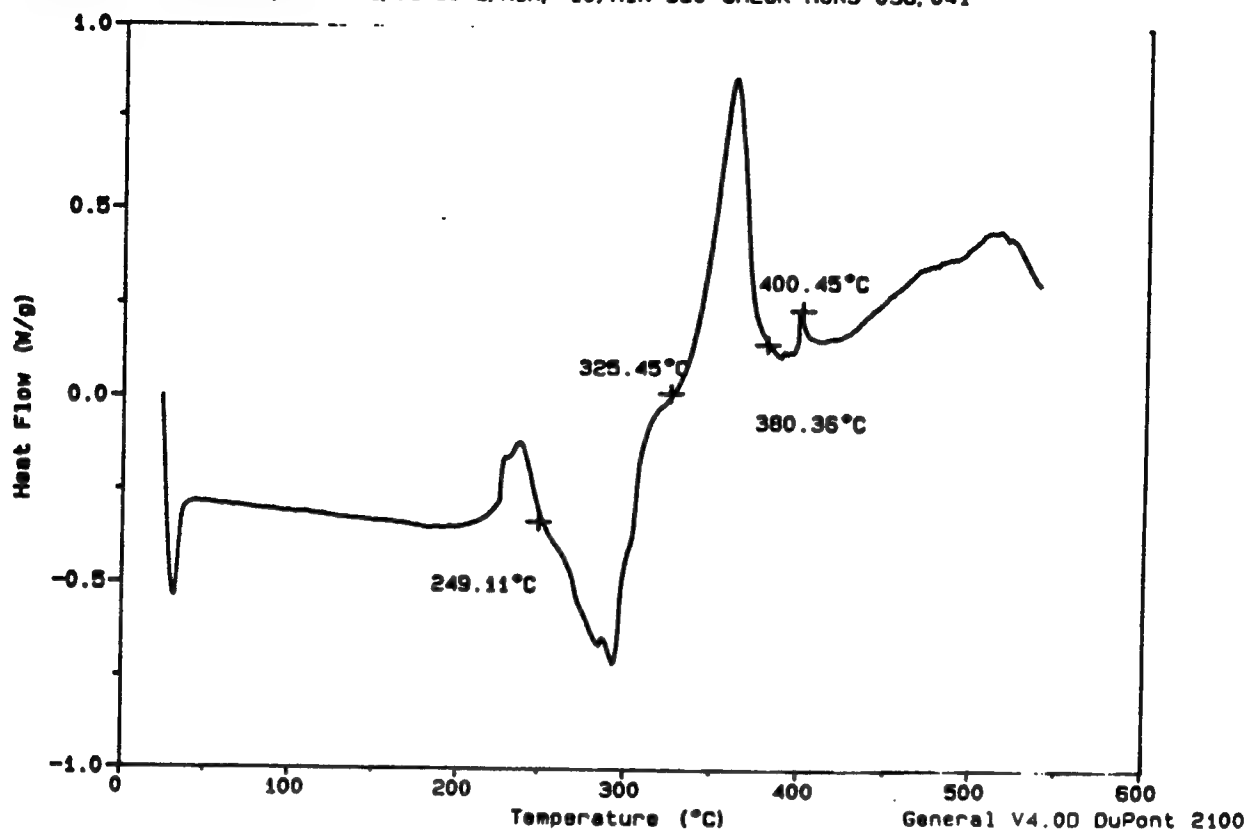


Fig.4 Sampling For HP-1050

The oxidized sample was dissolved with 1ml of THF(terahydrofuran), and 100µl of the THF solution was taken to inject into the HP-1050 LC to analyze the soluble organic compounds left in the sample. A summary of results is shown in the following table 1:

Table 1 : Non-Reacted TCP

Sample	Description	TCP Left
037	250° (11.18mg)	5.04µl(50.4%)
038	325° (11.18mg)	0.5µl(5%)
041	380° (11.35mg)	0
039	400° (11.5mg)	0

The HP-1050 LC printouts are also attached in figures 5, and 6.

Data File C:\HPCHEM\1\DATA\STT\DRLU037.D

Sample Name: DRLOC37

Sample 037, TCP 11.18 mg(10µl), 250°C, 1ml THF,  
100µl INJ VOL, 600µl GLASS VIAL.

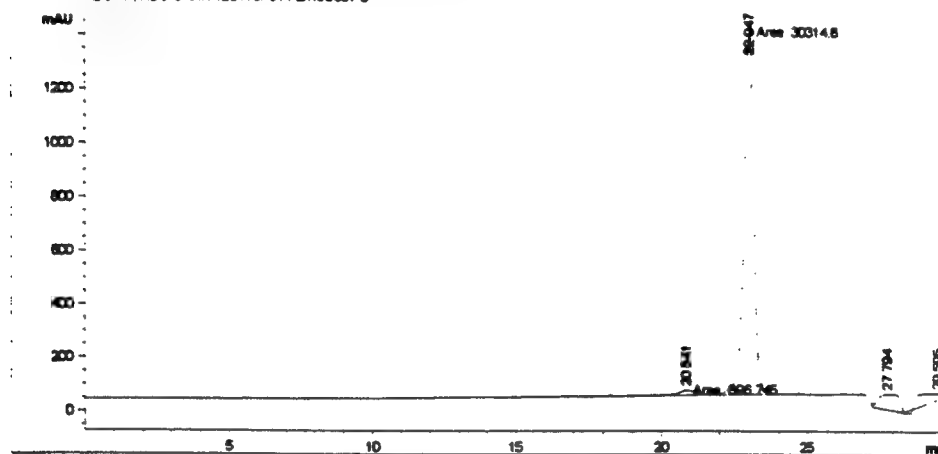
=====

Acq. Method	: HAKGPC.M	Seq. Line	: -
Acq. Operator	: WMM	Vial	: 12
Injection Date	: 6/20/95 1:03:53 PM	Inj	: -
Sample Name	: DRLU037	Inj Volume	: 100 µl

Analysis Method : C:\HPCHEM\1\METHODS\HAKGPC.M  
(modified after loading)

GPC analysis

ADC1 A, ADC CHANNEL A of STTDRLU037.D



Area Percent Report

Sorted by Signal  
Multiplier : 1.000000

Signal 1: ADC1 A, ADC CHANNEL A

Peak #	RT [min]	Type	Width [min]	Area	Height	Area %
1	20.841	MF	0.800	896.74451	18.67753	2.4706
2	22.947	PM	0.374	30314.83008	1350.17981	83.5191
3	27.794	PF	0.866	2941.91797	56.96931	8.1052
4	29.595	PBA	1.503	2143.38477	23.77052	5.9051

Totals : 36296.87891 1449.59717

Fig.5 HP Analysis, 250°C

Results from the LC analysis show that at 250°C, prior to the evaporation, the oxidation was already taking place, and a small amount of larger molecular weight compounds was formed at a retention time of 20.841 min in the figure 5. (Note: The retention time of the TCP molecules pass through the LC column is 22.947 min, and the larger the molecular weight the faster the

molecule passes through the column and the shorter the retention time.) Also almost 50% of TCP was still remained unreacted at 250°C prior to the evaporation (Table 1). At the end of evaporation, 325°C, only 5% of TCP was left unreacted (Table 1), and extra larger molecular weight compounds were formed at retention times of 19.985 min and 21.012 min (Fig.6).

Data File C:\HPCHEM\1\DATA\STT\DRLU038.D

Sample Name: DRLU038

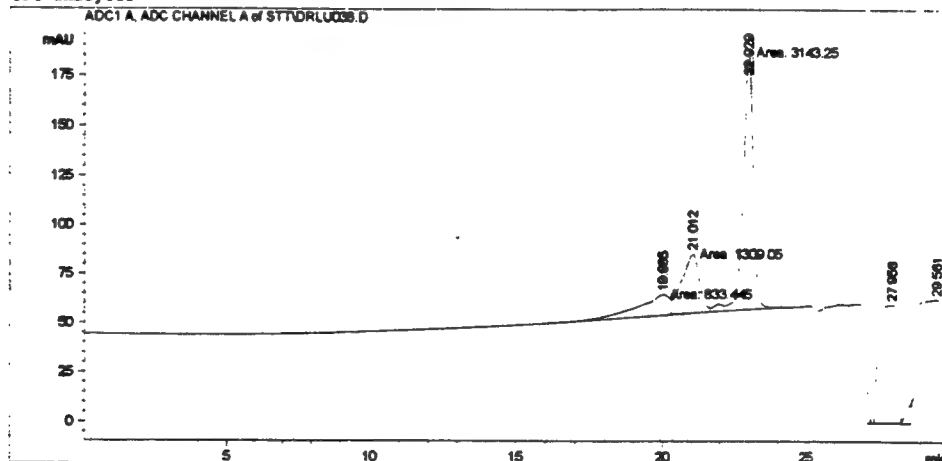
Sample 038, TCP 11.18 mg(10µl), 325°C, 1ml THF,  
100µl INJ VOL, 600µl GLASS VIAL.

=====

Acq. Method	: MAKGPC.M	Seq. Line	: -
Acq. Operator	: WNM	Vial	: 13
Injection Date	: 6/20/95 1:45:06 PM	Inj	: -
Sample Name	: DRLU038	Inj Volume	: 100 µl

Analysis Method : C:\HPCHEM\1\METHODS\MAKGPC.M  
(modified after loading)

GPC analysis



=====  
Area Percent Report  
=====

Sorted by Signal  
Multiplier : 1.000000

Signal 1: ADC1 A, ADC CHANNEL A

Peak #	RT [min]	Type	Width [min]	Area	Height	Area %
1	19.985	HF	1.292	833.44489	10.75413	8.1957
2	21.012	FM	0.720	1309.05444	30.31364	12.8726
3	22.929	FM	0.603	3143.25293	129.65775	30.9092
4	27.956	FF	0.645	2790.88452	59.85579	27.4442
5	29.561	FBA	1.274	2092.67871	27.37718	20.5784

Totals : 10169.31543 257.95850

HP LC System Thursday, June 29, 1995 3:12:35 PM by WNM

Page 1 of 1

Fig.6 HP Analysis, 325°C

No soluble organic compound was found at 380°C and 400°C samples at the end of a large exothermic reaction. At this stage, a combination of oxidation, polymerization, and possibly thermal decomposition may be taking place at a rather fast pace depending upon the amount of TCP originally started as well as the condition of heating rate inside the system. Our samples showed char residues were formed and they were insoluble with THF solvent.

## [II] Thermal Behavior of Iron

Iron powder was heated from the room temperature to 550°C at a rate of 10°C/min under both oxygen and nitrogen environments. For both cases and an exothermic peak was observed at above 550°C under the oxygen, and about 300°C to 350°C under the nitrogen environment (Fig. 7,8).

Sample: IRON (7.44MG) AIR 50ML/MIN

Size: 7.4400 mg

Method: 10°C/min 550°C, ISO 10 MN

Comment: OPEN AIR 10C/MIN TO 550C, HOLD 10 MIN

DSC

File: 95SUM.049

Operator: LU

Run Date: 29-Jun-95 09:44

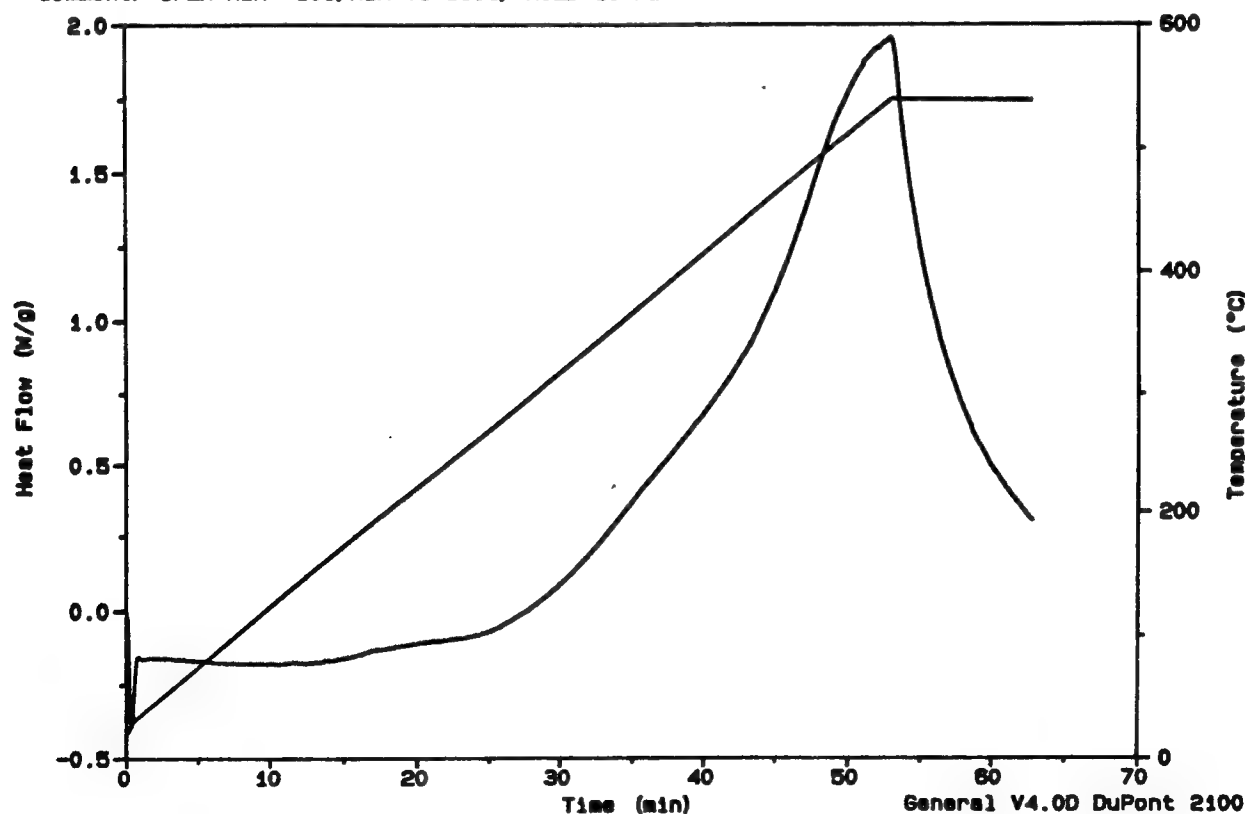


Fig.7 Iron Powder Under Air Flow

It is infeasible that iron will react with nitrogen, and these exothermic peaks are probably associated with the change of crystal structures from alpha or delta to gamma according to Cleaves and Thompson (Ref.6). It was discussed that alpha and delta iron have identical crystallographic structure of body-centered cubes, while gamma iron has a face-centered cubic structure. The alpha-gamma transformation takes place somewhat at 900°C. For a comparison, aluminum powder was also heated to 550°C with the air; the DSC thermograph showed a straight line without any signal of reaction. It is suggested that aluminum in any form, powder or sheet, is readily to form aluminum oxide, and aluminum oxide is stable in oxidation, noncombustible, insoluble in water, difficultly soluble in mineral acids and strong alkali.

Sample: IRON  
Size: 7.0800 mg

Method: 10°C/min to 550°C

Comment: OPEN N2 .50ML/MIN UL 10/MIN TO 550C, LOOK NO/OXIDATION OF FE

DSC

File: 95SUM.014

Operator: CCL

Run Date: 31-May-95 14:59

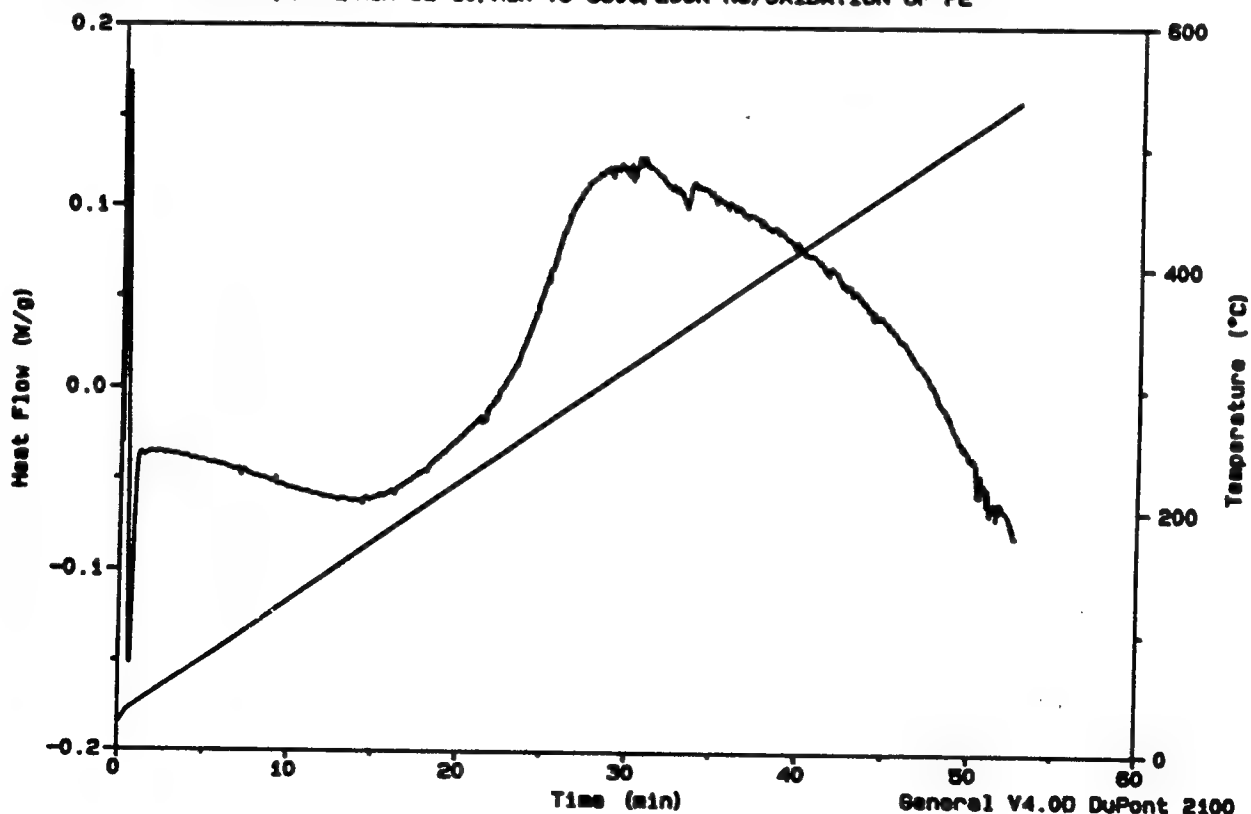


Fig.8 Iron Powder Under Nitrogen Flow

### [III] TCP/Iron Mixture

A typical TCP/iron mixture run is shown in figure 9, where 21.4mg of TCP was mixed with 5.2mg of iron powder, and was heated to 600°C under 50 ml/min of air flow.

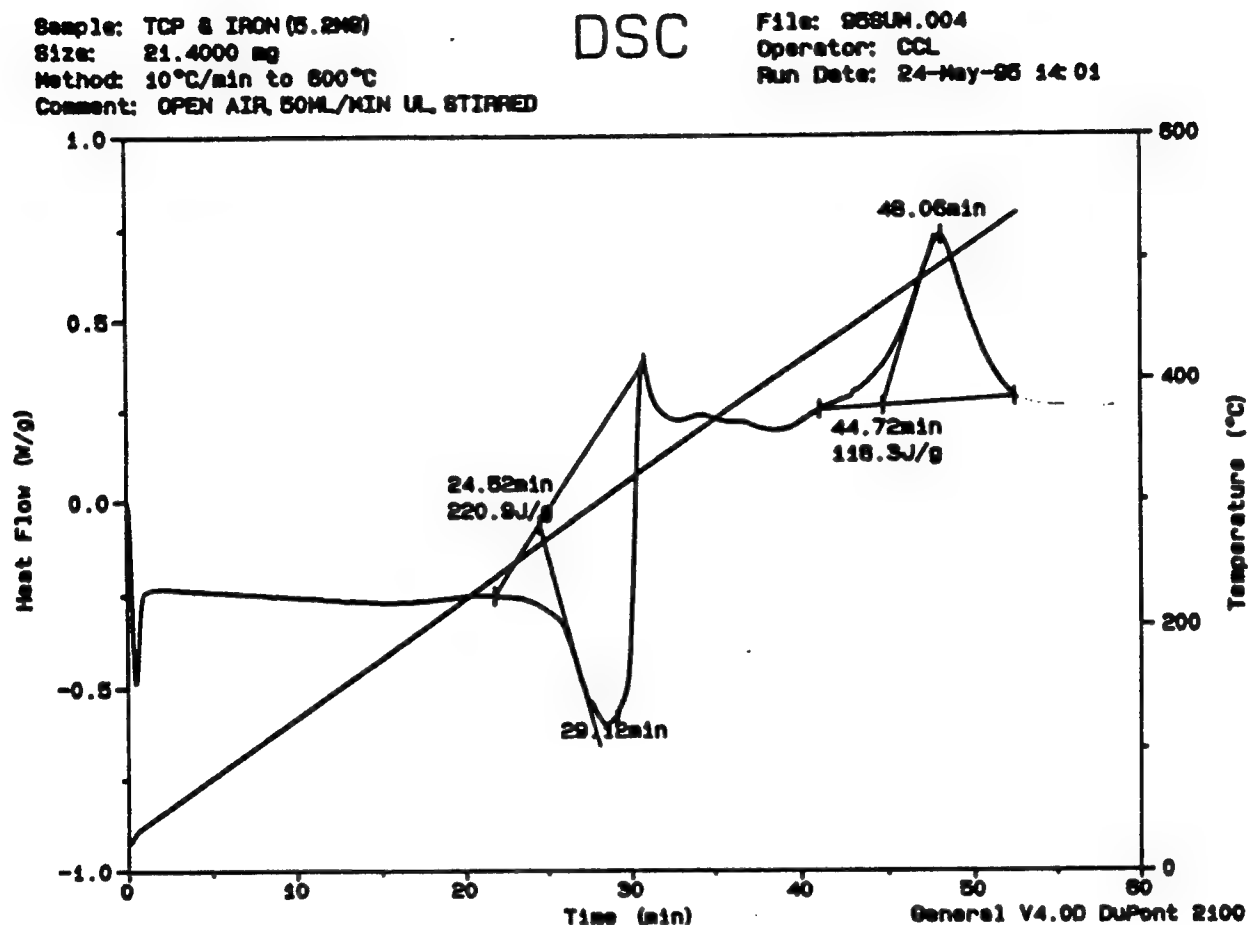


Fig.9 TCP & Iron Powder Under Open Air

An endothermic evaporation took place at about 320°C followed by some oxidation, and the last exothermic peak at about 500°C was possibly a combination of TCP char decomposition and iron crystal transformation. Some oxidized samples of the TCP/iron mixture were analyzed using SEM(scanning electron microscopy), a summary is shown in table 2. These samples were run at the air flow rate of 50 ml/min over aluminum pans without lids.



**Table 2: SEM TCP/Iron Mixture**

Sample	Descriptions	Comments
001	Blank TCP(18.2mg) 10°C/min to 550°C	P(major), Al round balls
003	Fe(4.0mg), TCP(43.2mg) 10°C/min to 550°C	Fe(large) Al(small) P(trace)
004	Fe(5.2mg), TCP(21.4mg) 10°C/min to 600°C	Fe & Al P(small) Si(very small)
005	Fe(3.36mg), TCP(25.4mg) 50°C Ramp to 400°C Hold 1 hr	Fe & Al P(medium) Si(small)
006	Fe(4.2mg), TCP(21.0mg) 50°C Ramp to 500°C Hold 1 hr	Fe(large) Al & P Mn(small)
007	Fe(3.4mg), TCP(24.5mg) 50°C Ramp to 400°C Hold 5 min	Fe(major) Al P(very small)
008	Fe(3.2mg), TCP(30.3mg) 50°C Ramp to 500°C Hold 7 min	Fe & Al(large) P(medium) Mn(small) Si(trace)

Photo copies of SEM are attached together in figure 10. In sample 001, two bright around spots indicate the existence of aluminum and phosphorus in the residue, phosphorus from TCP and aluminum from the pan. Other samples, 003 through 008, show different shapes of amorphous formations from the mixture of iron, phosphorus, and aluminum, etc. . . . The SEM actually can not identify whether these structures are crystal or amorphous. These residues were also dissolved with THF solution and analyzed using HP-1050 LC, the result showed no soluble organic compounds were left in the residue products.

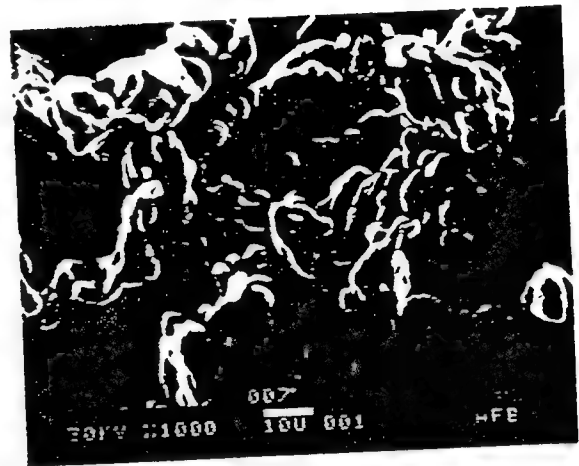
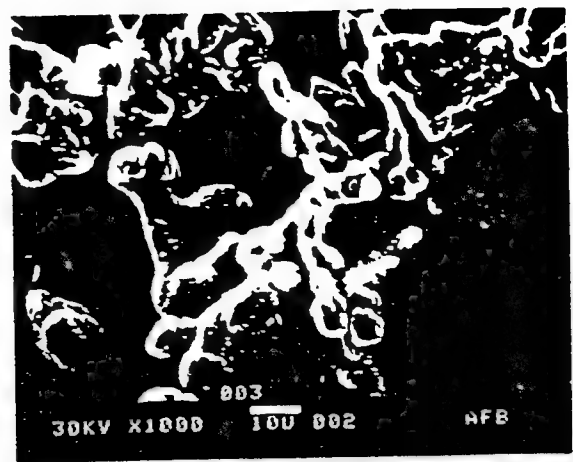
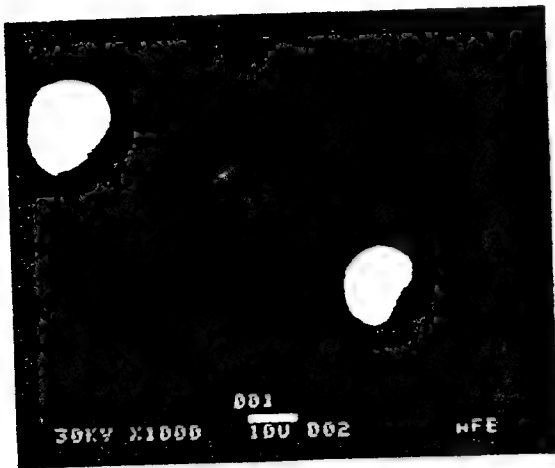


Fig.10 Photo Copies of SEM

## Conclusions and Recommendations

- 1). The DSC can be used to study the thermal behavior of liquid lubricants at various heating environments. Furthermore studies in the area of kinetics can be developed.
- 2). Using a mixture of liquid lubricant and metal powder, such as TCP/iron powder in this study, on the DSC to evaluate the reaction mechanism for a vapor phase lubrication is rather inappropriate, because most of liquid lubricants evaporate at about 300°C, and form chars before 350°C.

## Acknowledgements

The author thanks the Air Force Office of Scientific Research (AFSOR) 1995 Summer Research Program and the Lubrication Branch of Wright Patterson Air Force Base for this research opportunity. Thanks go to Michael A. Keller and Wesley M. Waldron, the University of Dayton Research Institute (UDRI), for technical discussions on this subject and analyzing data using HP-1050. Thanks also go to Christopher J. Klenke, WL/POSL, for SEM analysis. Special thanks go to Dr. Costandy S. Saba (UDRI) for using his facilities.

## References

1. Aldrich, Catalog Handbook of Fine Chemicals, 1992 - 1993.
2. CRC Handbook of Chemistry and Physics, 60th edition, CRC Press, Inc.
3. Lu, C. "Thermal Analysis and Molecular Weight Distribution of Triaryl Phosphates," Final Report for Summer Research Program, Lubrication Branch, Aero Propulsion and Power Directorate, Wright Laboratory (AFMC), Sept, 1992.
4. Lu, C. "Vaporization and Decomposition of Tricresyl Phosphate (TCP)," Final Report for Summer Research Program, WL/POSL, August, 1993.
5. Zhang, Y., Perez, P., and Hsu, S.M. "A New Method to Evaluate

Deposit-Forming Tendency of Liquid Lubricants by Differential Scanning Calorimetry," Lubrication Engineers, 48, 3, pp 189 - 1995, (1992).

6. Cleaves, and Thompson, "The Metal Iron, Chapter IV, Structure of High-Purity Iron".

**THE MEASUREMENT OF IMAGE POSITIONS  
IN CYLINDRICAL HOLOGRAMS.**

James S. Marsh  
Professor  
Department of Physics

The University of West Florida  
Pensacola, FL 32514

Final Report for:  
Summer Faculty Research Program  
Wright Laboratory  
Eglin AFB

Sponsored by:  
Air Force Office of Scientific Research  
Bolling Air Force Base, DC

and

Wright laboratory  
Eglin AFB

August 1995

**THE MEASUREMENT OF IMAGE POSITIONS  
IN CYLINDRICAL HOLOGRAMS**

James S. Marsh  
Professor  
Department of Physics  
The University of West Florida

Abstract

Using the tool of non-paraxial aberration theory developed by Champagne, a study is made of distortions and aberrations in images reconstructed from cylindrical holograms used in Ballistic Holography. The results are encapsulated in a suite of programs called FNDPNT and ABDIST, which display numerical and graphical results of calculations. The results show that it is generally possible to achieve resolutions of 100 microns or better, consistent with observations on the actual holograms. In general, distortion of the image field in comparison with the object field remains significant, even using 688 nm radiation from a laser diode to reconstruct a hologram made with 694 nm radiation from a ruby laser. However, the routines which form the basis for FNDPNT and ABDIST can be incorporated into measurement software to correct for this inherent distortion.

## **The measurement of image positions in cylindrical holograms.**

James S. Marsh  
The University of West Florida  
Pensacola, FL 32514

### **Introduction.**

The correlation of image positions with object positions in cylindrical holograms turns out to be unexpectedly complex. Unless the image is reconstructed with the same wavelength with which the hologram was made and the reconstruction beam emerges from the same place relative to the hologram as did the reference beam when the hologram was made, distortion of the image field relative to the object field in general occurs.

Since reconstruction of ballistic holograms made with 694 nm radiation from a pulsed ruby laser is, at best, practical with 688 nm radiation from a diode laser, this wavelength shift of slightly less than 1% must be taken into account for the highest precision measurements. The effect of this wavelength shift is, to a good approximation, to introduce a longitudinal magnification factor of the order of magnitude of 1% more or less than the lateral magnification factor in comparing the image field with respect to the object field. In addition, a small amount of shear along the axis of the hologram is introduced into the image field.

More significantly, it seems to be impractical to move the origin of the reconstruction beam from hologram to hologram to match up with the position of the reference beam in each case. Once the position of the origin of the reconstruction beam is fixed it is convenient to leave it there as long as possible due to the complexity of the optics associated with the laser diode. This means that the position of the reconstruction beam will differ significantly from that of the reference beam. This means that further distortion will be introduced into the image field, even if one could use 694 nm radiation in the reconstruction.

Related to this is the fact that the position of the reference beam is imperfectly known, only to within inches out of, typically, 40 inches for a 10 in. diameter hologram or 70 in. for an 18 in. hologram. In order to attain the highest accuracy of measurement relating image field positions to object field positions, the position of the reference beam as well as the reconstruction beam must be known to a suitably high accuracy. This means, given that the position of the reconstruction is well known, that the position of the reference beam must be determined from measurements of the positions of reference points in the image field and measurements of positions of the corresponding reference points in the object field.

Given the reconstruction wavelength, and the positions of the reference and reconstruction beams, the measured position of an image point turns out to depend on the piece of the hologram through which you are looking. This piece is called the pupil in this report. The pupil must therefore be known in order to relate image positions to object positions.

Finally, once you depart from the ideal reconstruction set up with the reconstruction wavelength identical to the reference wavelength and the reconstruction beam identical to the reference beam, aberrations are

introduced into the image. These must be estimated and minimized within the constraints of the optical set up.

The theory necessary to address these issues is called non-paraxial aberration theory. The "non"-paraxial part comes from the fact that the reconstruction and reference beams are generally making a large angle with respect to the normal of the hologram pupil. This necessitates a different expansion from the one used in ordinary aberration theory. Non-paraxial aberration theory was developed by Champagne<sup>1</sup> and further refined by Rebordao<sup>2</sup> and Peng and Frankena<sup>3</sup>.

In this report I will describe the results of non-paraxial aberration theory as applied to our situation of cylindrical holograms. Derivation of the theory is found in the references. I will also describe some programs I wrote in heavily commented Q-Basic that do some of the necessary calculations and use them to illustrate some of the points. Since the Basic language is practically plain English, the routines in these programs may be easily incorporated in what ever language is used for the analysis of the holograms.

#### Distortion.

Assume we are looking through a piece of the hologram, called the pupil. Denote by  $\mathbf{R}_r$  the vector pointing from the center of the pupil to the position of the reference beam, that is the point from which the spherical wave which is the reference beam originates,  $\mathbf{R}_c$  the vector to the reconstruction beam,  $\mathbf{R}_o$  the vector to the object point, and  $\mathbf{R}_i$  the vector from the pupil center to the image point. The magnitudes of these vectors will be denoted  $R_r$ ,  $R_c$ ,  $R_o$ , and  $R_i$ . The corresponding unit vectors will be denoted by  $\mathbf{R}_r$ ,  $\mathbf{R}_c$ ,  $\mathbf{R}_o$ , and  $\mathbf{R}_i$ .

Let the reference wavelength be  $\lambda_0$  and the reconstruction wavelength be  $\lambda$ . Then define

$$\mu = \lambda/\lambda_0.$$

Then the basic equations relating the various points relative to the pupil point are

$$\frac{1}{R_i} = \frac{1}{R_c} + \mu \left( \frac{1}{R_o} - \frac{1}{R_r} \right) \quad (1)$$

and

$$\mathbf{R}_i = \mathbf{R}_c + \mu (\mathbf{R}_o - \mathbf{R}_r) \big|_{\perp} \quad (2)$$

Equation (1) relates the magnitudes of the vectors to each other. Equation (2) relates the unit vectors. The notation  $\big|_{\perp}$  in equation (2) means that the components of these vectors in the plane of the pupil are related by the equation. Thus if the plane of the of pupil is the x-y plane then the x and y components of  $\mathbf{R}_i$  are determined by equation (2). The z component of  $\mathbf{R}_i$  is determined by the fact that  $\mathbf{R}_i$  is a unit vector. Then  $\mathbf{R}_i = R_i \mathbf{R}_i$ .

<sup>1</sup> Edwin B. Champagne, J. Opt. Soc. Am., 57, 51-55 (1967)

<sup>2</sup> J. M. Rebordao, J. Opt. Soc. Am. A, 1, 788-790 (1984)

<sup>3</sup> Ke-Ou Peng and Hans J. Frankena, Applied Optics, 25, 1319-1326 (1986)



Equations (1) and (2) are written in such a way that  $R_i$  is determined by knowledge of the other three  $R$ 's. But these equations may be used with any one of the  $R$ 's as unknown to be determined by the other three. These equations are used to determine the distortion of the object field in the reconstruction process and also to locate the position of the reference beam given the location of the reconstruction beam and the measured positions of an object point and a corresponding image.

### Aberrations.

In order to take a look at aberrations we must look at the whole wave field produced in the pupil of the hologram during the reconstruction. During recording of the hologram with wavelength  $\lambda_0$ , an object point produces a wave of the form  $\exp(ikr_0)$  in the pupil of the hologram, where  $r_0$  is the distance from the object point to the place in the pupil being inspected; in vector terms  $r_0 = -R_0 + r$ ,  $R_0$  the vector from the center of the pupil to the object point as defined above, and  $r$  points from the center of the pupil to the point of the pupil being inspected.  $k$  is the recording wave number,  $k=2\pi/\lambda_0$ .

$r_r$  and  $r_c$  are defined similarly to  $r_0$ ;  $r_r = -R_r + r$  and  $r_c = -R_c + r$ . If the hologram is illuminated in the reconstruction with light of wavelength  $\lambda$ , then in the pupil of the hologram a wave of the form

$$A = \exp(ik'r_c)\exp(ikr_0)\exp(-ikr_r)$$

is produced, where  $k'$  is the wavenumber of the reconstruction beam,  $k'=2\pi/\lambda$ . We write this as

$$A = \exp(ik'\{r_c + \mu(r_0 - r_r)\})$$

where  $\mu = \lambda/\lambda_0$  as defined previously.

Now  $A$  must be of the form

$$A = \exp(ik'r_i)\exp(i\Delta\phi)$$

where  $r_i = -R_i + r$  points from the image point to the place in the pupil being inspected.  $\Delta\phi$  is the phase error and forms the aberrations of the image point. We rewrite the two expressions for  $A$  to get

$$\exp(i\Delta\phi) = \exp(ik'(-r_i + r_c + \mu(r_0 - r_r))). \quad (3)$$

The expressions on the right are expanded in powers of  $r$ . Requiring the terms with the lowest order in  $r$  to be constant leads to equations (1) and (2), which locate  $R_i$  in terms of the other  $R$ 's. The higher order terms in  $r$  give the various wave aberrations in the image.

We determine the smallest resolvable point at an image point by choosing  $|r|$  and numerically calculating the wave aberration,  $-r_i + r_c + \mu(r_0 - r_r)$ , at  $5^\circ$  intervals of  $r$ . We choose the smallest value of  $|r|$  at which the wave aberration exceeds  $\lambda/4$  to be the optimum pupil size for that pupil and image point. The idea is that if the pupil was

smaller than the optimum pupil size, the size of the image of the point would increase due to diffraction effects, while if the pupil was larger than the optimum size, the image size would increase due to aberrations. The smallest resolvable point size is then determined to be  $d = \lambda / \text{N.A.}$ .

N.A. is the numerical aperture,  $\text{N.A.} = \rho / \sqrt{\rho^2 + R_i^2}$ , where  $\rho$  is the optimum value of  $|\mathbf{r}|$ .

#### The Programs.

I have written two programs in QBasic, FINDPNT and ABDIST. ABDIST comes in three versions, ABDISTYZ2, ABDISTXZ2, and ABDISTXY2. Both programs are based on the following geometry.

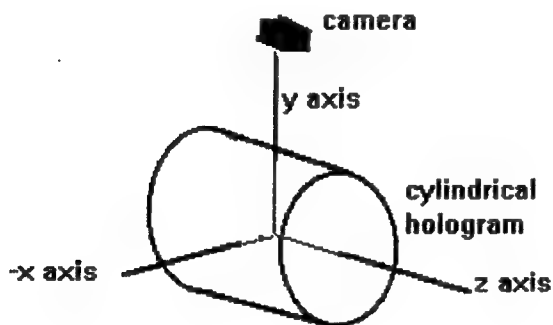


Figure 1a

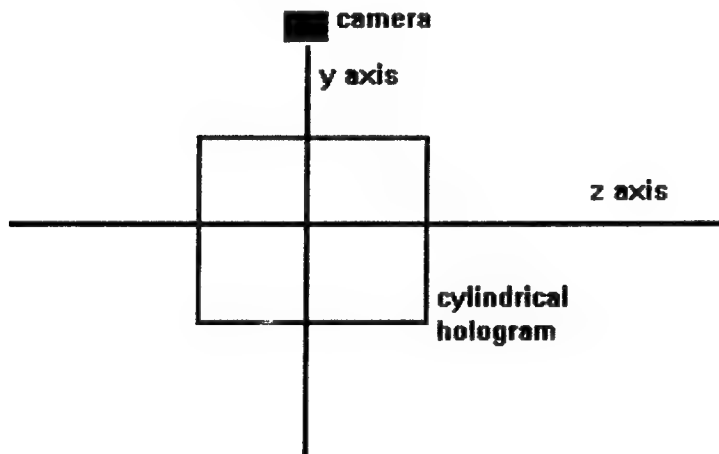


Figure 1b

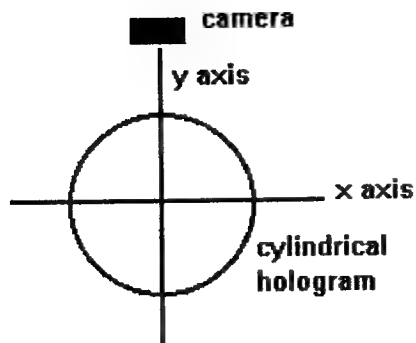


Figure 1c

Figure 1a is a 3-dimensional representation of the geometry while Figure 1b is a yz plane projection and Figure 1c is an xy projection of the geometry. The origin of the coordinate system is in the center of the hologram. The z axis coincides with the axis of the hologram. The y axis extends up toward the camera. All positions of points are input to or returned by the programs in component form relative to these axes.

#### FINDPNT.

FINDPNT stands for FIND the PoiNT. It uses equations (1) and (2) to find any one of the points after the input of the other three. The position of the pupil point must be put in by hand.

The program first asks for the reconstruction wavelength in nm. It is assumed that the recording wavelength is 694 nm.

Then it asks for the radius of the hologram. This may be in any units, inches, cm., mm, etc. Then all subsequent coordinates are assumed to be in the same units.

Now it asks for the coordinates of the pupil. Since the pupil must be on the hologram itself, the program asks for the x and the z coordinate of the pupil. The program calculates the y coordinate of the pupil and returns all three.

The program then asks which point is to be calculated from the other three. The coordinates of the points are input and returned in the form x,y,z.

As an example, suppose a reconstruction at 632 nm is done on a 5 in. radius hologram. Suppose the position of an image point is measured and it is desired to find where the corresponding object point is located. The reference beam was at 40 in. from the end of the hologram on the hologram axis. This makes the coordinates of the reference beam relative to the origin at the middle of the hologram  $x_r = 0$ ,  $y_r = 0$ ,  $z_r = 45$ . The extra 5 inches comes from the fact that the hologram width is 10 inches from side to side.

Suppose the reconstruction beam is 30 in. from the end of the hologram and 2 in. off axis in the -y direction, i.e. opposite to the direction where the camera sits. This makes the coordinates of the reconstruction beam  $x_c = 0$ ,  $y_c = -2$ ,  $z_c = 35$ .

Suppose the image point is measured by the camera to be on the negative y axis at  $y=-4$  in. Then the coordinates of the image point are  $x_i=0, y_i=-4, z_i=0$ .

Assuming the camera was on the positive y axis, this means the pupil will be on the y axis. The input for the pupil coordinates will be  $x_p=0$  and  $z_p=0$ . The program will return  $x_p=0, y_p=5, z_p=0$ .

The output from FINDPNT will be as in the next figure.

```
Find the missing point.
reconstruction wavelength (nm)? 632
cylinder radius(any units)? 5

Enter coordinates of pupil.
xp= 0          yp= 5          zp= 0

Selected point is Object point.

position of Reference point.
xr 0          yr 0          zr 45

position of reConstruction point.
xc= 0          yc=-2          zc= 35

position of Image point.
xi= 0          yi=-4          zi= 0

Position of Object point.
xo= 0          yo=-3.793227  zo=-.731415
```

**Press any key to continue**

We see that the object point whose image position was measured to be  $x_i=0, y_i=-4, z_i=0$ , was actually located at  $x_o=0, y_o=-3.79, z_o=-.73$ . So the image point position was shifted by  $-.21$  in. the y direction and  $.73$  in. toward the reference beam in the reconstruction.

One of the uses of this program will be to find out where the reference beam was from the position of the reconstruction beam and measured positions of a fiducial object point and the corresponding image point. This measured position of the reference point can then be used in subsequent data analysis to correlate image point positions with corresponding object point positions.

As an example, we had a hologram with a  $12\text{cm.} = 120\text{ mm}$  radius. The reconstruction beam was  $33\text{ in.} = 825\text{ mm}$  on axis from the origin. The position of the far end of the actual cylinder was at  $x_o=0, y_o=-120\text{ mm}, z_o=3\text{ in.} = 75\text{ mm}$ . The image of that point was measured to be at  $x_i=0, y_i=-115\text{ mm.}, z_i=3\text{ in.}=75\text{ mm}$ . The reconstruction was done at  $688\text{ nm}$ . It was assumed that the pupil point was at  $x_p=0, z_p=0$ . FINDPNT returned the results

Find the missing point.  
reconstruction wavelength (nm)? 688  
cylinder radius(any units)? 120

Enter coordinates of pupil.  
xp= 0            yp= 120            zp= 0

Selected point is Reference point.

position of reConstruction point.  
xc= 0            yc= 0            zc= 825

position of Image point.  
xi= 0            yi=-115            zi= 75

position of Object point.  
xo= 0            yo=-120            zo= 75

Position of Reference point.  
xr= 0            yr=-9.780151    zr= 901.6731

**Press any key to continue**

The reference beam was therefore 902mm=36in. up the z axis from the origin or 33 in from the end of the cylinder. Furthermore it was off axis -9.9mm. = -.4 in. in the y direction.

One of the uses of FINDPNT is to explore quantitatively the consequences of equations (1) and (2) regarding distortion and shifting about of the image location for a given object point as the pupil changes. The dependence of image location on pupil location was mentioned in the introduction.

I used FINDPNT to examine this last question. Suppose we reconstruct at 688 nm a 5 in. radius hologram with an object point at the origin at the center of the hologram,  $x_o=0, y_o=0, z_o=0$ . Suppose we managed to make the reconstruction beam coincide with the reference beam at  $x_r=x_c=0, y_r=y_c=0, z_r=z_c=30$  in. Putting the pupil point at  $x_p=0$ , and exploring the image position for different positions,  $z_p$ , of the pupil point along the z axis yielded the following results.

zp	xi	yi	zi
-5	0	.166	-.186
-4	0	.161	-.139
-3	0	.156	-.097
-2	0	.153	-.059
-1	0	.153	-.023
0	0	.159	.013
1	0	.170	.054
2	0	.189	.104
3	0	.215	.169
4	0	.250	.254
5	0	.294	.364

Here we see different amounts of y and z displacement of the image point relative to the object point for different positions of the pupil. This is even though in other respects the reconstruction geometry is nearly ideal with the reconstruction beam identical with the reference beam and the reconstruction wavelength at 688 nm compared to 694 nm for the recording wavelength. These displacements of the image point from the object point are as much as 5 mm and never less than 3.8 mm in the y direction and range from over +4 to -4 mm in the z direction. This means that even with laser diode radiation in the reconstruction corrections to the measured positions of images will be necessary to locate the corresponding object points to better than 3 or 4 mm. Since image positions can be determined to an order of magnitude better precision than this and, in addition, 100 micron resolution or better is achieved by our measuring apparatus, calculation of object locations from measured image locations is a desirable thing to do. The routines used in FINDPNT and ABDIST can be adapted to work with the measurement routines.

#### ABDIST.

ABDIST stands for ABerrations and DISTortions. The ABDIST programs pick a plane within the holographic volume and calculate the positions of the corresponding object points for a regular array of image points in the chosen plane. Thus the distortion field of the given plane may be viewed. In addition, if the aberration option is selected, each image point is decorated with a circle whose size indicates the smallest resolvable point at that position. Thus the resolution field can also be displayed.

The geometry on which ABDIST is based is the same as that for FINDPNT as shown figures 1. ABDISTYZ2 shows the distortion and aberration field for the YZ plane, ABDISTXZ2 for the xz plane, and ABDISTXY2 for the xy plane.

ABDIST asks for the reconstruction wavelength in nm and the radius of the cylinder. As currently set up, the cylinder radius and the coordinates of the various points are expected in inches.

ABDIST asks for positions of the reference and reconstruction beam in the same format as FINDPNT. However, when requesting the position of the reconstruction beam, ABDIST offers the option of choosing the reconstruction beam to come in at the Bragg angle for the middle of the hologram receiving the radiation from an object point at the origin of the coordinates. If this option is chosen, ABDIST asks for  $R_c$ , the distance of the reconstruction beam from the pupil in the center of the hologram.

In addition, ABDIST will ask for the coordinates of the camera, more explicitly the entrance pupil of the camera. With this information ABDIST will calculate the location of the pupil on the hologram for the various image points.

Finally, ABDIST offers the option to look at the aberration field. If this option is chosen it asks for a resolution scale. The number 100 is a typical value to choose for the resolution scale. If this number is chosen the program will tell you that one pixel = 10 microns. This means that a circle of radius one pixel is drawn about the image point that a particle size resolution of 10 microns is possible at that point. If the radius of the circle is 10 pixels, then a particle size resolution of 100 microns is possible, and so on. Smaller resolution scale numbers will show coarser resolutions while larger resolution scale number will show finer resolutions.

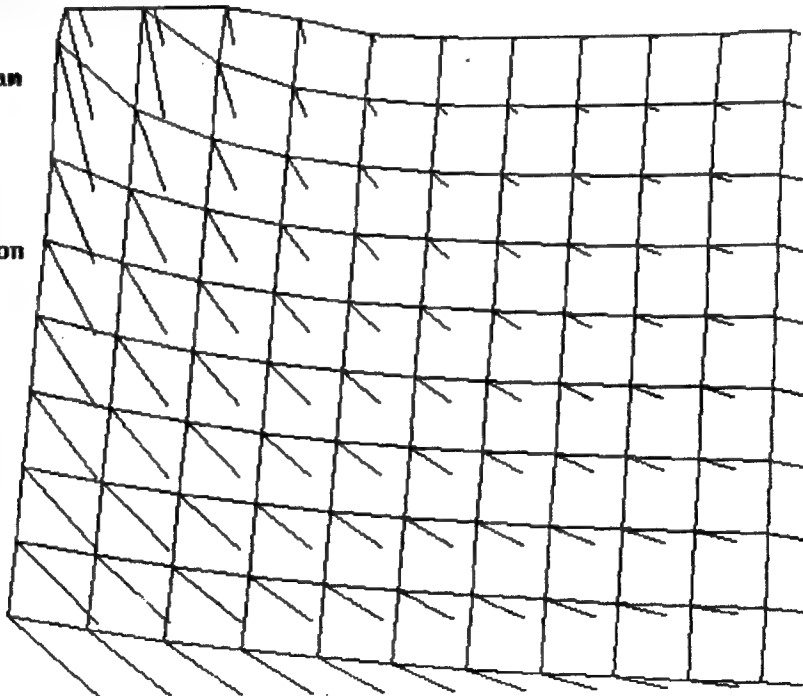
As an example, suppose we have a 5 in. radius hologram illuminated with 632 nm light from 35 in. away from the origin on the z axis. Assume the reference beam was 40 in. away from the origin on the z axis. First we will take the camera to be just at the middle of the hologram, 1 in. away from the hologram. We will have  $x_v=0, y_v=6, z_v=0$ . The subscript v stands for video. Ignoring the resolution option we get, from ABDISTYZ2,

yz plane distortion and aberrations.  
reconstruction wavelength (nm)? 632  
cylinder radius(in)? 5

position of  
reference beam  
xr? 0  
yr? 0  
zr? 40

position of  
reconstruction  
beam  
xc? 0  
yc? 0  
zc? 35

position of  
video camera  
xv? 0  
yv? 6  
zv? 0



Press any key to continue

In this figure the one ended lines begin with their dangling ends on image points. The image points are in a square array from -R to R along the z axis (which points to the right in the figure) in 11 steps and from -R+.5 to R-.5 along the y axis (which points up in the figure) in 10

steps. The single ended lines end on corresponding object points. The object points are joined together with a mesh of lines which show the distortion of the square mesh of image points to the distorted mesh of corresponding object points.

This case shows extreme distortion of the object field due to the fact that the camera is so close to the hologram it must swivel through a large arc to sweep out the whole image field. From study of this figure and others like it we determined that distortion will be much less over the whole holographic volume if the camera looks straight into the hologram rather than swiveling. This is achieved by translating the camera rather than swiveling it to sweep out the holographic volume.

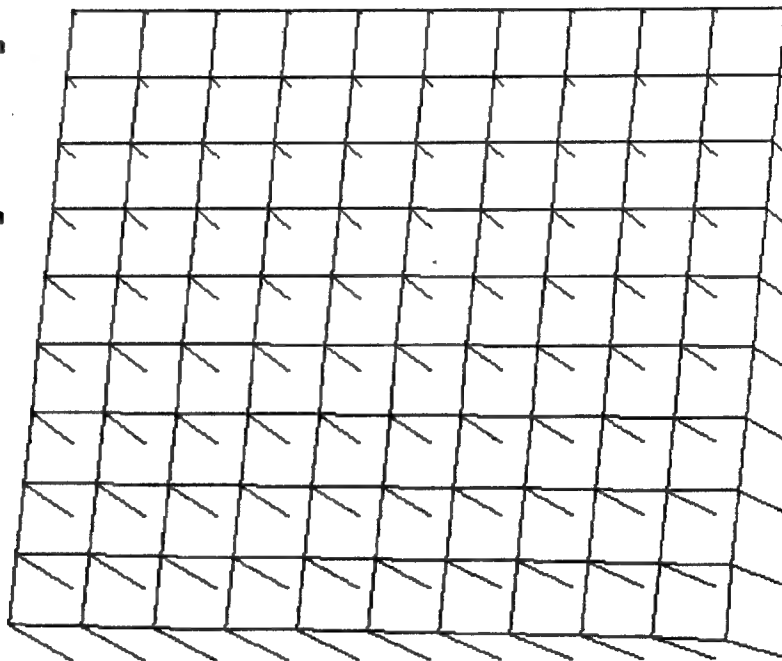
This is simulated imperfectly in the next figure when the camera is placed far away from hologram<sup>4</sup>, everything else left the same.  
yv=1000 in.

yz plane distortion and aberrations.  
reconstruction wavelength (nm)? 632  
cylinder radius(in)? 5

position of  
reference beam  
xr? 0  
yr? 0  
zr? 40

position of  
reconstruction  
beam  
xc? 0  
yc? 0  
zc? 35

position of  
video camera  
xu? 0  
yu? 1000  
zu? 0



Press any key to continue

Now we see that the effect is much more regular over the holographic volume. The lines of the object mesh are more nearly straight lines. The effect can be reproduced, to a good approximation, with a y direction magnification factor of some value greater than one, and a z axis shearing factor. Numerical values for these factors could be easily derived from FINDPNT.

Now let's do this again but now show the aberration field.

<sup>4</sup> In practice the camera needs to be close to the hologram but translated to sweep out the holographic volume. The program needs to be modified to simulate this.



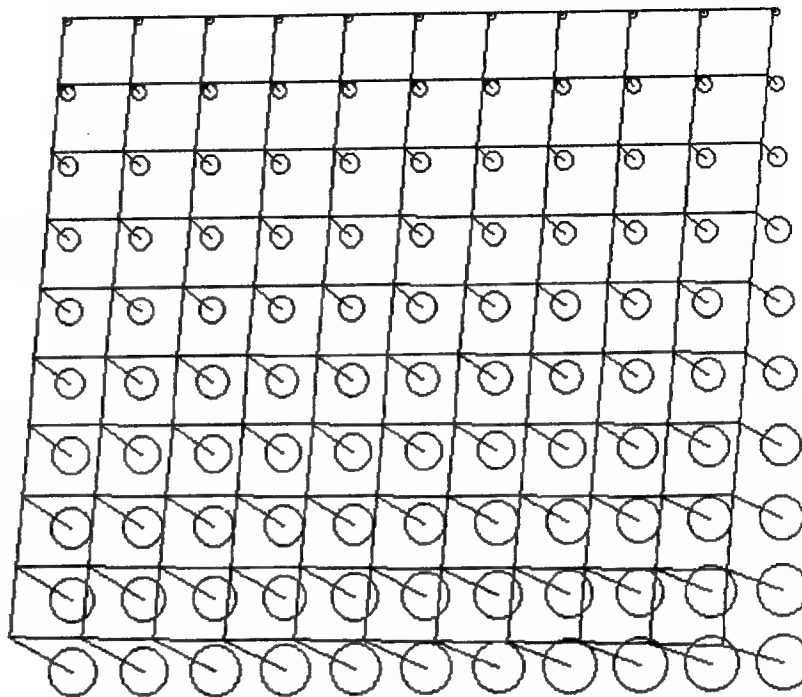
yz plane distortion and aberrations.  
reconstruction wavelength (nm)? 632  
cylinder radius(in)? 5

position of  
reference beam  
xr? 0  
yr? 0  
zr? 40

position of  
reconstruction  
beam  
xc? 0  
yc? 0  
zc? 35

position of  
video camera  
xu? 0  
yu? 1000  
zv? 0

resolution  
scale? 100  
1 pixel =  
10 microns



Press any key to continue

This shows that the resolution is, at worst, about 150 microns under these conditions.

Now we look at the same thing but view the xz plane using ABDISTXZ2.

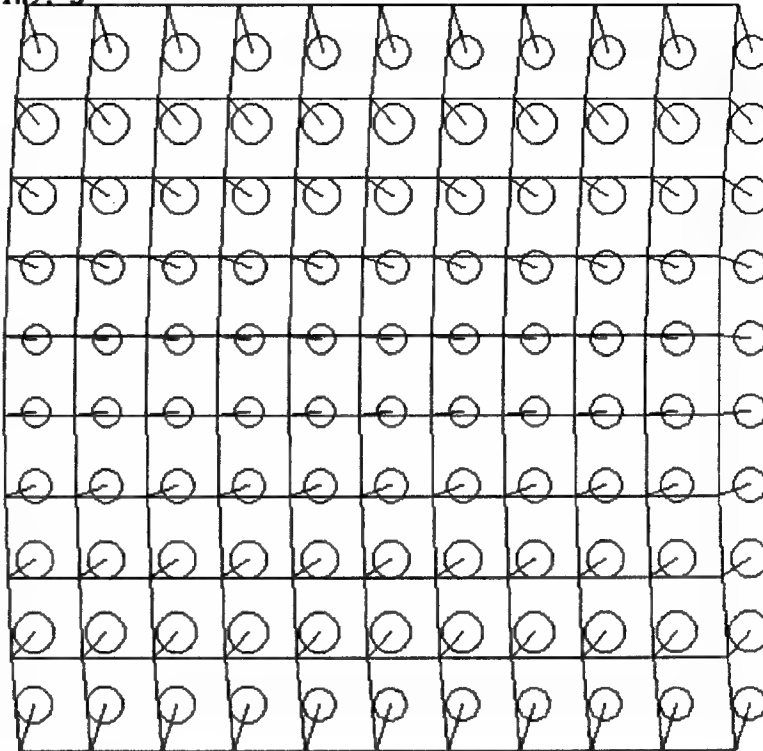
xz plane distortion and aberrations.  
 reconstruction wavelength (nm)? 632  
 cylinder radius(in)? 5

position of  
 reference beam  
 xr? 0  
 yr? 0  
 zr? 40

position of  
 reconstruction  
 beam  
 xc? 0  
 yc? 0  
 zc? 35

position of  
 video camera  
 xv? 0  
 yv? 1000  
 zv? 0

resolution  
 scale? 100  
 1 pixel =  
 10 microns



Press any key to continue

In this figure, the z axis is toward the right and the x axis points up. The image points are at the centers of the circles while the corresponding object points are joined by the mesh. The displacements are projected down onto the xz plane. Each of the displacements has a component perpendicular to the xz plane, some of which can be seen in the previous figure. Bowing of some of the mesh lines is apparent. This, however, is not significant if image point-object point correspondences based on equations (1) and (2) are incorporated into the measuring software. More significant is the fact that resolution in the xz planes hovers around the 100 micron value.

Finally we look at the xy plane under these conditions using ABDISTXY2.

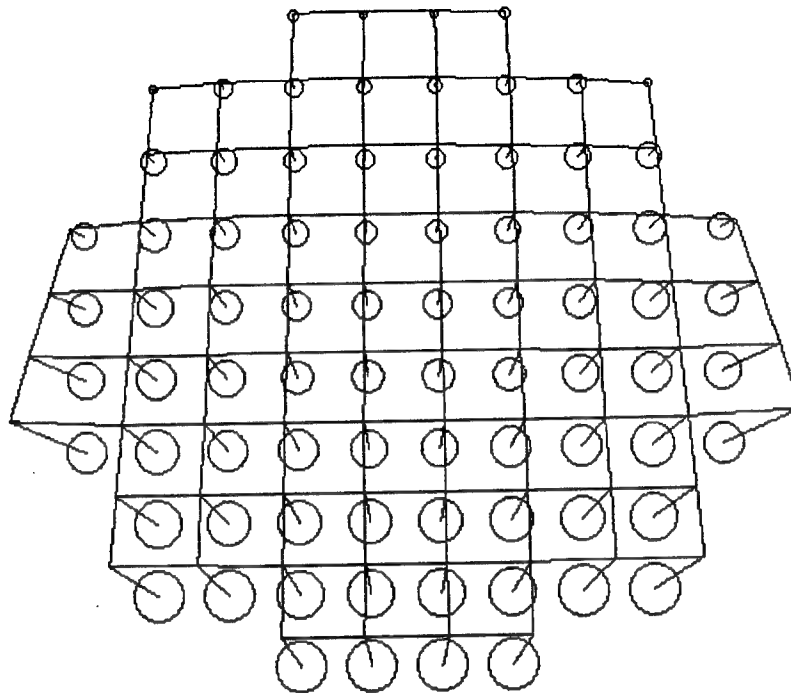
xy plane distortion and aberrations.  
reconstruction wavelength (nm)? 632  
cylinder radius(in)? 5

position of  
reference beam  
xr? 0  
yr? 0  
zr? 40

position of  
reconstruction  
beam  
xc? 0  
yc? 0  
zc? 35

position of  
video camera  
xv? 0  
yv? 1000  
zv? 0

resolution  
scale? 100  
1 pixel =  
10 microns



Press any key to continue

Here we see considerable distortion out toward the sides of the hologram. Resolution remains better than 150 microns over all. The lesson from these figures is that, if 150 micron resolution at worst is acceptable, then reconstruction with 632 radiation is perfectly acceptable since the distortions can be calculated away.

Now we will look at a similar set of charts of reconstructions using 688 radiation, but with everything else the same.

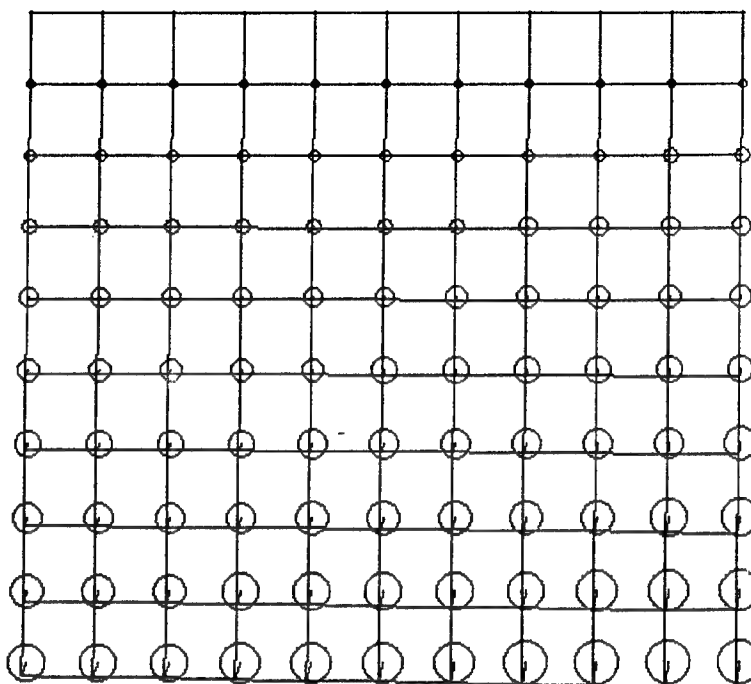
yz plane distortion and aberrations.  
 reconstruction wavelength (nm)? 688  
 cylinder radius(in)? 5

position of  
 reference beam  
 xr? 0  
 yr? 0  
 zr? 40

position of  
 reconstruction  
 beam  
 xc? 0  
 yc? 0  
 zc? 35

position of  
 video camera  
 xv? 0  
 yv? 1000  
 zv? 0

resolution  
 scale? 100  
 1 pixel =  
 10 microns



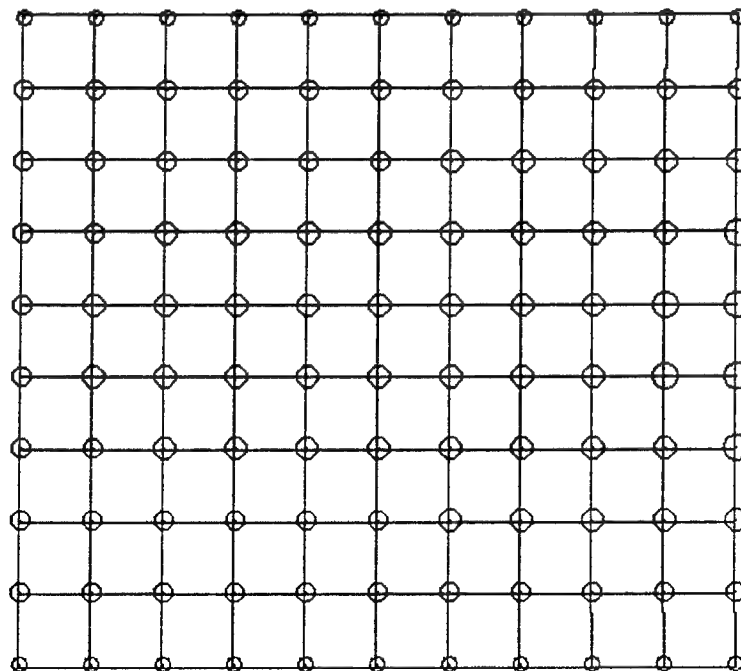
Press any key to continue  
 xz plane distortion and aberrations.  
 reconstruction wavelength (nm)? 688  
 cylinder radius(in)? 5

position of  
 reference beam  
 xr? 0  
 yr? 0  
 zr? 40

position of  
 reconstruction  
 beam  
 xc? 0  
 yc? 0  
 zc? 35

position of  
 video camera  
 xv? 0  
 yv? 1000  
 zv? 0

resolution  
 scale? 100  
 1 pixel =  
 10 microns



Press any key to continue

xy plane distortion and aberrations.  
reconstruction wavelength (nm)? 688  
cylinder radius(in)? 5

position of  
reference beam

xr? 0  
yr? 0  
zr? 40

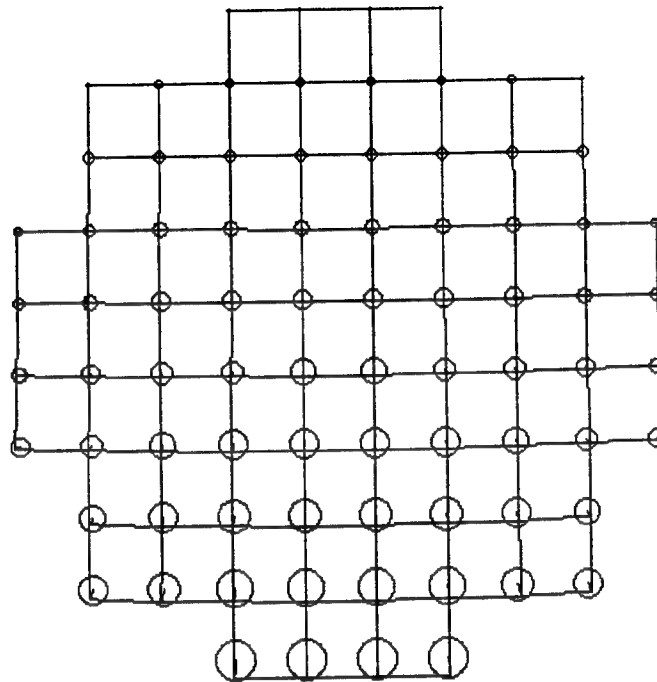
position of  
reconstruction  
beam

xc? 0  
yc? 0  
zc? 35

position of  
video camera

xv? 0  
yv? 1000  
zv? 0

resolution  
scale? 100  
1 pixel =  
10 microns



Press any key to continue

In general we see that under the same conditions of reference beam and reconstruction beam, the 688 nm reconstruction gives better resolution than the 632 nm reconstruction. However, in *both* cases, reconstruction conditions can be found which will improve the resolution. What is significant is that the displacement of object points from image points is significant even for 688 nm radiation. This means that the necessary distortion corrections must be made in the measurement software.

COMPUTATIONAL STUDIES OF THE REACTIONS OF ATOMIC HYDROGEN WITH  
FLUOROMETHANES: KINETICS AND BRANCHING RATIOS

Paul Marshall  
Associate Professor  
Department of Chemistry

University of North Texas  
PO Box 5068, Denton, Texas 76203-0068

Final Report for:  
Summer Faculty Research Program  
Wright Laboratory, Materials Directorate, Wright-Patterson AFB, Ohio 45433

Sponsored by:  
Air Force Office of Scientific Research  
Bolling AFB, Washington, D.C.

and

Wright Laboratory

September 1995

# COMPUTATIONAL STUDIES OF THE REACTIONS OF ATOMIC HYDROGEN WITH FLUOROMETHANES: KINETICS AND BRANCHING RATIOS

Paul Marshall  
Associate Professor  
Department of Chemistry  
University of North Texas  
PO Box 5068, Denton, Texas 76203-0068

## Abstract

Geometries and vibrational frequencies for transition states in the reactions of H with CH<sub>4</sub>, CH<sub>3</sub>F, CH<sub>2</sub>F<sub>2</sub>, CHF<sub>3</sub> and CF<sub>4</sub> have been characterized at the MP2(FU)/6-31G(d) level of theory. The Gaussian-2 methodology yielded barrier heights at the approximate QCISD(T)/6-311+G(3df,2p) level of calculation. The results were employed to calculate rate constants via transition state theory with a tunneling correction. The results are in good accord with experimental rate constants, where these are available for comparison. H-abstraction, F-abstraction and F-substitution pathways were considered, and the results show that, contrary to some earlier assumptions, H atoms react predominantly with the C-H bonds in fluoromethanes and that the major product channel is H<sub>2</sub> production. F atom abstraction is unfavorable kinetically, even though HF formation is the most exothermic pathway. For CF<sub>4</sub> this is the only possible pathway for H-atom attack, which is therefore slow under combustion conditions so that CF<sub>4</sub> is essentially inert in a flame. Thus CF<sub>4</sub> acts mainly as a physical flame suppressant. By contrast, the other fluoromethanes react quickly with H atoms to yield F-containing radicals that undergo further chemistry, which opens the possibility of chemical flame suppression by CH<sub>3</sub>F, CH<sub>2</sub>F<sub>2</sub> and CHF<sub>3</sub>.

# COMPUTATIONAL STUDIES OF THE REACTIONS OF ATOMIC HYDROGEN WITH FLUOROMETHANES: KINETICS AND BRANCHING RATIOS

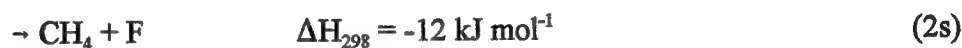
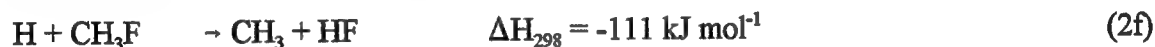
Paul Marshall

## INTRODUCTION

Kinetic data for the reactions of H atoms with fluorinated methanes are needed to model the combustion chemistry of fluorine-containing compounds. This topic has received recent attention through efforts to understand the flame suppression activity of halons such as CF<sub>3</sub>Br and potential substitutes.<sup>1,2,3,4</sup> However, by contrast to the process



the experimental kinetic data base for fluorinated molecules is sparse and sometimes contradictory. For example, three pathways for attack by H on CH<sub>3</sub>F have been discussed in the literature,<sup>5</sup> abstraction of F “f”, abstraction of H “h”, and substitution “s”



The exothermicities<sup>6,7</sup> demonstrate that all these channels are thermochemically reasonable. The dominance of the F-abstraction channel 2f is generally accepted or assumed in experimental work,<sup>5</sup> but Westmoreland et al.<sup>4</sup> have recently argued that H-abstraction 2h is the most important pathway, based on BAC-MP4 calculations. At the same time estimates of the total rate constant  $k_2$  differ by more than 2 orders magnitude at 600 K, where different measurement techniques overlap.<sup>5</sup>

Here the results of high-level ab initio calculations are presented for abstraction and substitution pathways for the series  $\text{H} + \text{CH}_{4-x}\text{F}_x$  ( $x=0-4$ ):





Barrier heights were estimated using Gaussian-2 theory.<sup>8</sup> G2 theory approximates QCISD(T)/6-311+G(3df,2p) and yields accord with experimental atomization energy for bound molecules with an average absolute deviation of 5 kJ mol<sup>-1</sup>. One aim of the present work is to see whether a similar degree of accuracy holds for various transition states. The G2 results are employed here in transition state theory calculations to yield rate expressions and product branching ratios for reactions 1-5, and the implications for models of flame suppression by fluorinated agents are discussed.

## METHODOLOGY

The geometries of the reactants, transition state (TS) and products for each reaction were initially optimized at the HF/6-31G(d) level of theory. For each species vibrational frequencies were scaled by a standard factor (0.8929) before calculation of the zero-point energy (ZPE). Next, the geometry was refined at the MP2(FU)/6-31G(d) level of theory, and vibrational frequencies again obtained. These verified that true TS geometries were derived, with single complex frequencies that correspond to motion along the reaction coordinate. Finally, approximate QCISD(T)/6-311+G(3df,2p) energies were obtained via the G2 protocol, as detailed elsewhere,<sup>8</sup> through a series of additive corrections to the MP4/6-311G(d,p) energy. These calculations were carried out by means of the Gaussian 92 program.<sup>9</sup>

The bimolecular rate constant *k* for each pathway was derived by application of canonical transition state theory (TST) implemented with the Polyrate 6.5 program.<sup>10</sup>

$$k_{\text{TST}} = \Gamma \frac{k_B T}{h} \frac{Q_{\text{TS}^\ddagger}}{Q_{\text{H}} Q_{\text{CH}_4-x\text{F}_x}} \exp\left(-\frac{E_0^\ddagger}{RT}\right) \quad (6)$$

$\Gamma$  represents a correction factor for quantum mechanical tunneling derived at the zero-curvature level and is  $\geq 1$ .<sup>11</sup> The partition functions  $Q$  include rotational symmetry numbers, and  $E_0^\ddagger$  is the energy difference between reactants and TS including ZPE, i.e., the enthalpy difference at 0 K.  $k$  for the reverse reaction is obtained by microscopic reversibility, since the ratio of forward and reverse rate constants is the equilibrium constant  $K_{\text{eq}}$ .

## RESULTS AND DISCUSSION

The calculated structures of the TSs are shown in Fig. 1, the vibrational frequencies in Table 1, and the components of the G2 energies in Table 2. It can be seen that for the series  $\text{CH}_4 - \text{CHF}_3$ , the barrier  $E_0^\ddagger$  does *not* vary monotonically with the number of C-F bonds, as might be expected, but does correlate with the calculated TS geometries: the higher barriers are for TSs with longer breaking C-H bonds and shorter forming H-H bonds. The Evans-Polanyi plot shown in Fig. 2 indicates that knowledge of the reaction enthalpy does have predictive power for the barrier to H-abstraction, although the trend of the G2  $E_0^\ddagger$  values as a function of the G2  $\Delta H_0$  is not strictly monotonic.

Comparison with previous analysis of TS properties is possible only for reaction 1h, and Table 3 overleaf shows some representative earlier calculations of the barrier for  $\text{H} + \text{CH}_4$ .<sup>12,13,14,15,16,17</sup> The value calculated here is seen to be in good accord with value derived from fitting to experimental data. An Arrhenius plot of the TST rate constant  $k_{1h}$  is presented in Fig. 3, where calculated results are compared to experimental values.<sup>18</sup> It may be seen that there is considerable scatter among

**Table 3:** Barrier heights for the reaction  $\text{H} + \text{CH}_4 \rightarrow \text{CH}_3 + \text{H}_2$  calculated at various levels of theory.

$E_0^\ddagger$ , kJ mol <sup>-1</sup>	method	ref.
65.1	POL-CI	12
59.4	PMP4SDTQ	13
55.8	CCSD(T)	14
57.0	QCISD(T)	15
61.9	QCISD	16
62.7	Fit to expt.	17
60.1	G2	this work

the measurements and that our TST calculations support the smaller  $k_{1h}$  values. These are probably the more reliable measurements, because unrecognized secondary chemistry will lead to overestimated rate constants. This potential experimental problem becomes acute when the rate constant is small, i.e., below around  $10^{-15} \text{ cm}^3 \text{ molecule}^{-1} \text{ s}^{-1}$ .

The barriers to F-atom abstraction increase along the series  $\text{CH}_3\text{F}-\text{CF}_4$  and there is a good correlation between the G2  $E_0^\ddagger$  and  $\Delta H_0$  values, illustrated on Fig. 2, although the MP2 C-F and H-F distances in the TSs do not vary consistently. Because the barriers to abstraction of H atoms are significantly lower than for F-atom abstraction we expect the former channel to dominate under all conditions. TSs for substitution of F by H have also been characterized: the higher barrier for  $\text{CH}_2\text{F}_2$  as compared to  $\text{CH}_3\text{F}$  suggests the barriers would be even higher for  $\text{CHF}_3$  and  $\text{CF}_4$ .

Fig. 4 illustrates the computed rate constants for H reactions with fluoromethane. The dominant product channel is H-atom abstraction, 2h, and the best accord between experiment and theory is for a set of measurements where knowledge of the products was not necessary for

interpretation of the observations.<sup>19</sup> The minor channels 2f and 2s are predicted to have similar, small, rate constants. For reaction 3,  $\text{H} + \text{CH}_2\text{F}_2$ , H abstraction is again predicted to be more favorable (see Fig. 5), although there are no experimental data for comparison. Figure 6 shows that for  $\text{H} + \text{CHF}_3$  there is good accord between the measured rate constant and  $k_{\text{th}}$ , and that F abstraction is unfavorable. For  $\text{H} + \text{CF}_4$ , F abstraction is the only channel: the rate constant  $k_{\text{sf}}$  for this slow process is in reasonable agreement with experiment (Fig. 7).

## CONCLUSIONS

Transition state theory based on ab initio Gaussian-2 data gives good accord with experimental rate constants, where these are available for comparison. The results show that, contrary to some earlier assumptions, H atoms react predominantly with the C-H bonds in fluoromethanes and that the major product channel is  $\text{H}_2$  production. F atom abstraction is unfavorable kinetically, even though HF formation is the most exothermic pathway. For  $\text{CF}_4$  this is the only possible pathway for H-atom attack, which is therefore slow under combustion conditions so that  $\text{CF}_4$  is essentially inert in a flame. Thus  $\text{CF}_4$  acts mainly as a physical flame suppressant. By contrast, the other fluoromethanes react quickly with H atoms to yield F-containing radicals that undergo further chemistry, which opens the possibility of chemical flame suppression by  $\text{CH}_3\text{F}$ ,  $\text{CH}_2\text{F}_2$  and  $\text{CHF}_3$ .

## REFERENCES

1. Westbrook, C. K.; *Combust. Sci. Tech.* **1983**, *34*, 201.
2. Battin-Leclerc, F.; Côme, G. M.; Baronnet, F. *Comb. Flame* **1994**, *99*, 644.
3. Richter, H.; Vandooren, J.; van Tiggelen, P. J. *25th Symp. (Int.) Combust.* (The Combustion Institute, Pittsburgh, 1994) p. 825.
4. Westmoreland, P. R.; Burgess, Jr., D. R. F.; Tsang, W.; Zachariah, M. R. *25th Symp. (Int.) Combust.* (The Combustion Institute, Pittsburgh, 1994) p. 1505.
5. Baulch, D. L.; Duxbury, J.; Grant, S. J.; Montague, D. C. "Evaluated Kinetic Data for High Temperature Reactions. Vol. 4. Homogeneous Gas Phase Reactions of Halogen- and Cyanide-Containing Species", *J. Phys. Chem. Ref. Data* **1981**, *10*, Suppl. 1.
6. Chase Jr., M. W.; Davies, C. A.; Downey, Jr., J. R.; Frurip, D. J.; McDonald, R. A.; Syverud, A. N. *JANAF Thermochemical Tables*, 3rd Ed., *J. Phys. Chem. Ref. Data* **1985**, *14*, Suppl. 1.
7. McMillen, D. F.; Golden, D. M. *Ann. Rev. Phys. Chem.* **1982**, *33*, 493.
8. Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. *J. Chem. Phys.* **1991**, *94*, 7221.
9. Frisch, M. J.; Trucks, G. W.; Head-Gordon, V.; Gill, P. M. W.; Wong, M. W.; Foresman, J. B.; Johnson, B. G.; Schlegel, H. B.; Robb, M. A.; Replogle, E. S.; Gomperts, V.; Andres, J. L.; Raghavachari, K.; Binkley, J. S.; Gonzalez, V.; Martin, R. L.; Fox, D. J.; DeFrees, D. J.; Baker, J.; Stewart, J. J. P.; Pople, J. A. *GAUSSIAN92* (Gaussian, Pittsburgh, 1992).
10. Steckler, R.; Hu, W.-P.; Liu, Y.-P.; Lynch, G. C.; Garrett, B. C.; Isaacson, A. D.; Lu, D.-h.; Melissas, V. S.; Truong, T. N.; Rai, S. N.; Hancock, G. C.; Lauderdale, J. G.; Joseph, T.; Truhlar, D. G. *POLYRATE* - version 6.5, University of Minnesota, Minneapolis, 1995.
11. Garrett, B. C.; Truhlar, D. G.; Grev, R. S.; Magnuson, A. W. *J. Phys. Chem.* **1980**, *84*, 1730; **1983**, *87*, 4454 (E).

12. Walch, S. P. *J. Chem. Phys.* **1980**, *72*, 4932.
13. Gonzalez, C.; McDouall, J. J. W.; Schlegel, H. B. *J. Phys. Chem.* **1990**, *94*, 7467.
14. Kraka, E.; Gauss, J.; Cremer, D. *J. Chem. Phys.* **1993**, *99*, 5306.
15. Dobbs, K. D.; Dixon, D. A. *J. Phys. Chem.* **1994**, *98*, 5290.
16. Truong, T. N. *J. Chem. Phys.* **1994**, *100*, 8014.
17. Marquaire, P.-M.; Dastidar, A. G.; Manthorne, K. C.; Pacey, P. D. *Can. J. Chem.* **1994**, *72*, 600.
18. Mallard, W. G.; Westley, F.; Herron, J. T.; Hampson, R. F. *NIST Chemical Kinetics Database*, Ver. 6.0 NIST Standard Reference Data, Gaithersburg, MD (1994).
19. Westenberg, A. A.; De Haas, N. J. *J. Chem. Phys.* **1975**, *62*, 3321.

**Table 1:** Vibrational frequencies for transition states (cm<sup>-1</sup>)

Reactants	Transition state	HF/6-31G(d) <sup>a</sup>	MP2(FU)/6-31G(d) <sup>b</sup>
H + CH <sub>4</sub>	TS1h	2003i, 545(2), 1094, 1165(2), 1363, 1409(2), 2901, 3022(2)	1856i, 589(2), 1130, 1222 (2), 1500 (2), 1851, 3169, 3327(2)
H + CH <sub>3</sub> F	TS2h	2094i, 294, 533, 1076, 1114, 1161, 1213, 1228, 1364, 1457, 2933, 3025	2059i, 310, 579, 1138, 1174, 1229, 1305, 1311, 1546, 1707, 3164, 3282
	TS2f	2021i, 417(2), 650, 739(2), 1110, 1418(2), 2943, 3080(2)	2803i, 368(2), 854(2), 1044, 1264, 1517(2), 3192, 3354(2)
	TS2s	1757i, 458(2), 535, 1153(2), 1223, 1366(2), 2931, 3091(2)	1742i, 362(2), 889, 1358(2), 1430(2), 1589, 3110, 3291(2)
H + CH <sub>2</sub> F <sub>2</sub>	TS3h	2144i, 276, 352, 521, 1056, 1135, 1148, 1198, 1254, 1381, 1390, 2990	2118i, 292, 372, 539, 1120, 1191, 1200, 1265, 1363, 1442, 1697, 3188
	TS3f	2155i, 237, 436, 494, 701, 723, 1121, 1148, 1179, 1476, 2988, 3103	2873i, 274, 418, 492, 822, 1076, 1198, 1246, 1250, 1570, 3212, 3349
	TS3s	2044i, 303, 470, 508, 814, 1104, 1118, 1231, 1249, 1409, 2952, 3086	1631i, 277, 398, 803, 1063, 1119, 1222, 1422, 1482, 1741, 2865, 3204
H + CHF <sub>3</sub>	TS4h	2147i, 271(2), 496(2), 662, 1005, 1146(2), 1301(2), 1441	2036i, 282(2), 509(2), 688, 1051, 1208(2), 1344(2), 1815
	TS4f	2220i, 211, 294, 482, 516, 582, 744, 1122, 1155, 1210, 1381, 3049	2863i, 242, 337, 494, 518, 610, 1106, 1195, 1227, 1260, 1444, 3244
H + CF <sub>4</sub>	TS5f	2172i, 208 (2), 480 (2), 503, 554, 782, 1075, 1327(2)	2774i, 236(2), 485(2), 565(2), 605, 960, 1255, 1357(2)

<sup>a</sup>Scaled by 0.8929<sup>b</sup>Unscaled

**Table 2:** Absolute G2 energies of transition states for H + fluoromethanes calculated at the MP2(FU)/6-31G(d) optimized geometries.

Species	Sym.	State	MP4/6- 311G(d,p) <sup>a</sup>	$\langle S^2 \rangle^b$	$\Delta E(+)^c$	$\Delta E(2df)^c$	$\Delta E(QCD)^c$	$\Delta E(ZPE)^c$	$\Delta^c$	$E(G2)^a$	$E_0^{\dagger d}$
TS1h	C <sub>3v</sub>		-40.87725	0.787	-0.56	-18.71	-2.94	+40.19	-8.55	-40.88801	60.1
TS2h	C <sub>s</sub>	<sup>2</sup> A'	-139.94181	0.789	-8.93	-69.76	-1.88	+35.08	-11.74	-140.03422	52.5
TS2f	C <sub>3v</sub>		-139.90695	0.840	-14.91	-68.54	-4.51	+36.48	-11.90	-140.00553	127.8
TS2s	C <sub>3v</sub>		-139.90625	0.833	-13.49	-70.99	-4.91	+38.33	-12.41	-140.00490	129.4
TS3h	C <sub>s</sub>	<sup>2</sup> A'	-239.02565	0.789	-14.21	-121.91	-0.30	+28.94	-15.16	-239.19848	51.2
TS3f			-238.98063	0.844	-21.26	-119.89	-3.41	+31.00	-15.11	-239.15948	153.5
TS3s			-238.96729	0.812	-18.72	-122.00	-3.85	+32.45	-16.28	-239.14588	189.2
TS4h	C <sub>3v</sub>		-338.11683	0.789	-17.69	-174.94	+1.50	+21.73	-18.73	-338.37015	61.5
TS4f			-338.06732	0.846	-25.21	-172.58	-2.25	+24.48	-18.28	-338.32635	176.5
TS5f	C <sub>3v</sub>		-437.15510	0.846	-27.38	-226.04	-0.94	+17.08	-21.60	-437.49416	189.4

<sup>a</sup>In au. 1 au  $\approx$  2625 kJ mol<sup>-1</sup>.

<sup>b</sup>For the HF/6-311G(d,p) wavefunction.

<sup>c</sup>Component of G2 energy in 10<sup>-3</sup> au.

<sup>d</sup>G2 enthalpy relative to H + fluoromethane at 0 K, in kJ mol<sup>-1</sup>.



Fig. 1. MP2(FU)/6-31G(d) geometries for transition states.

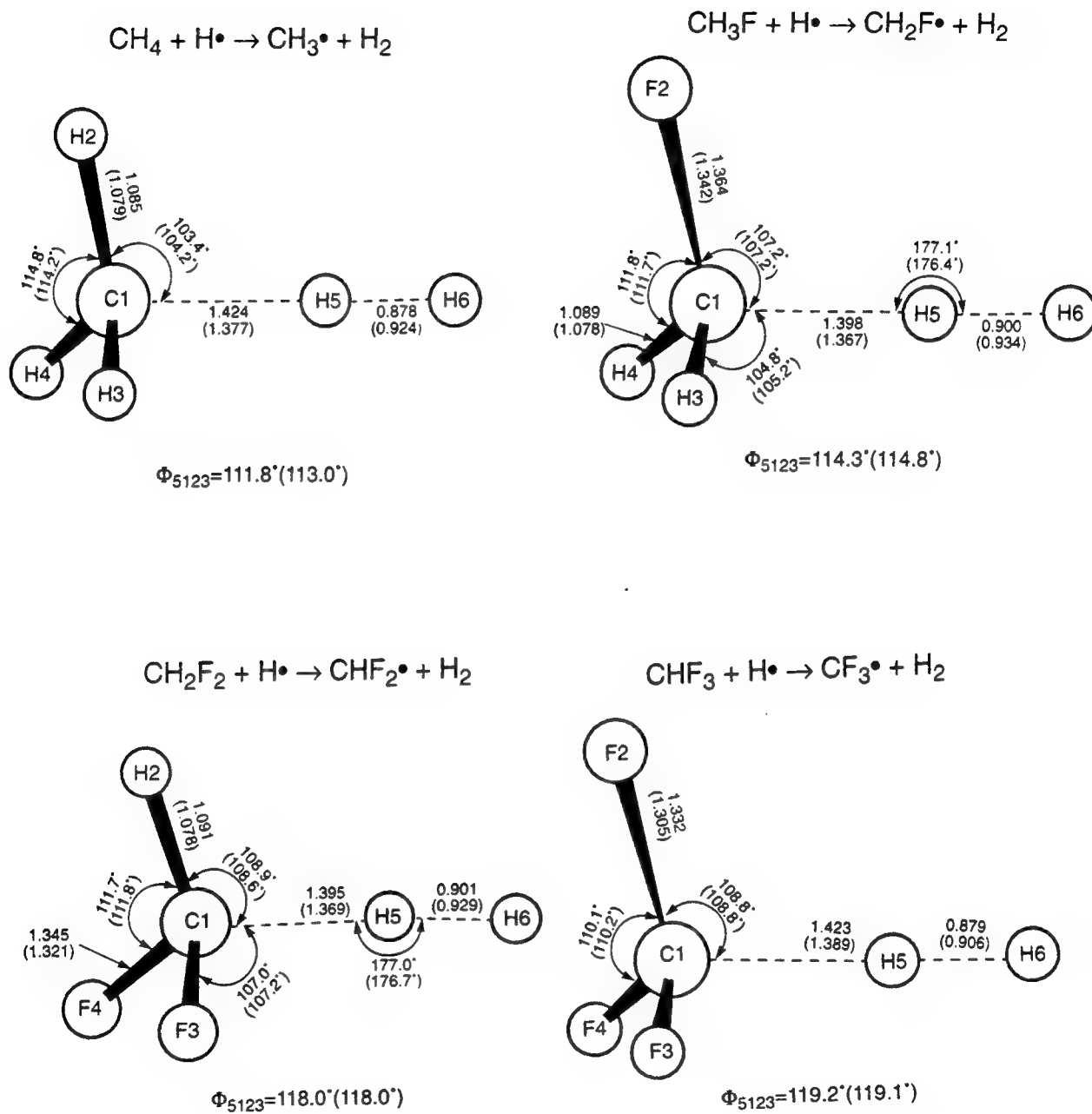


Fig. 1. Contd.

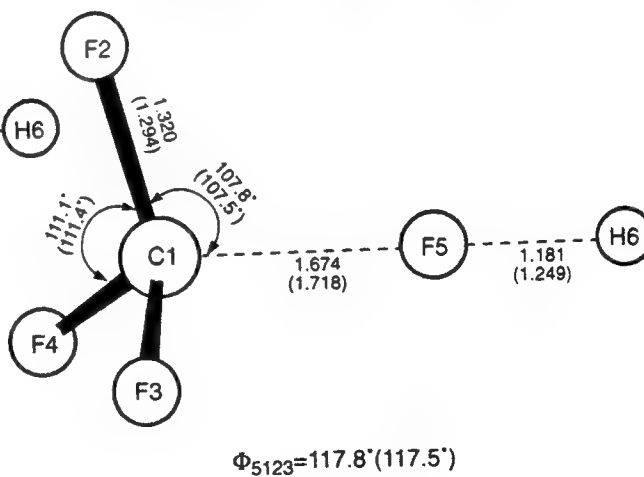
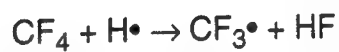
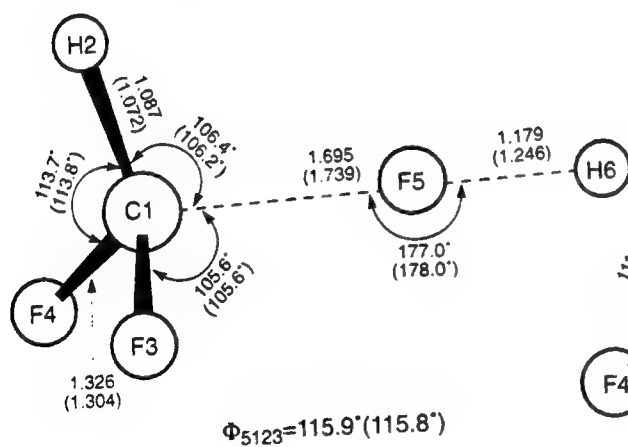
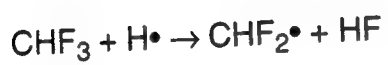
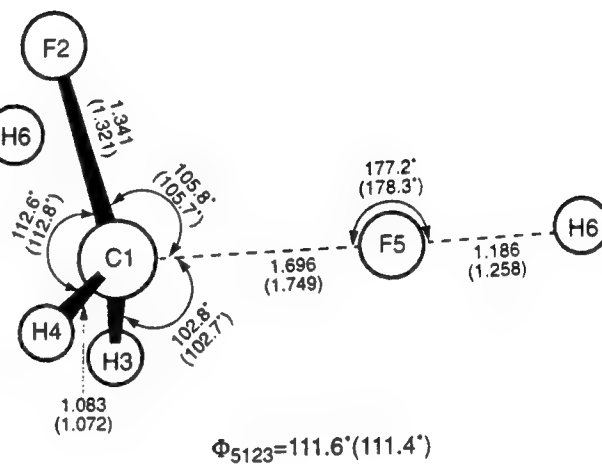
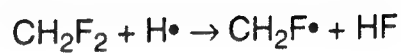
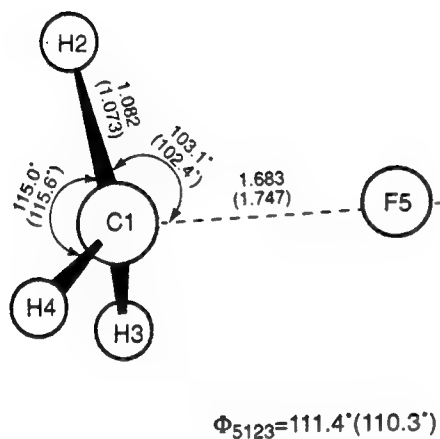
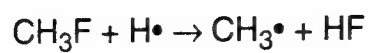


Fig. 1. Contd.

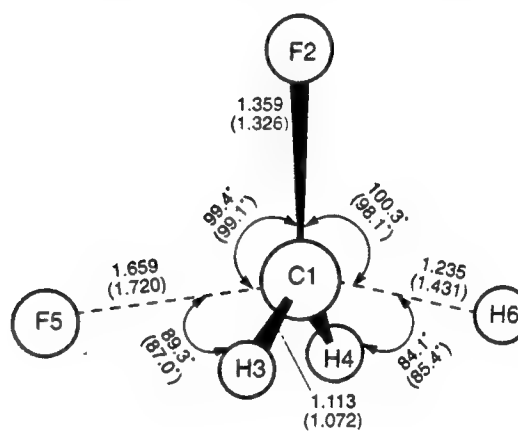
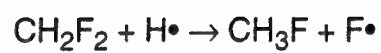
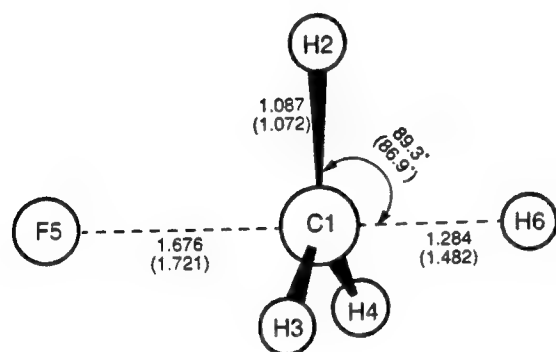
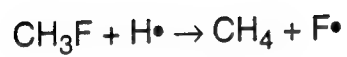


Fig. 2. Evans-Polanyi plot for reactions of H atoms with fluoromethanes.

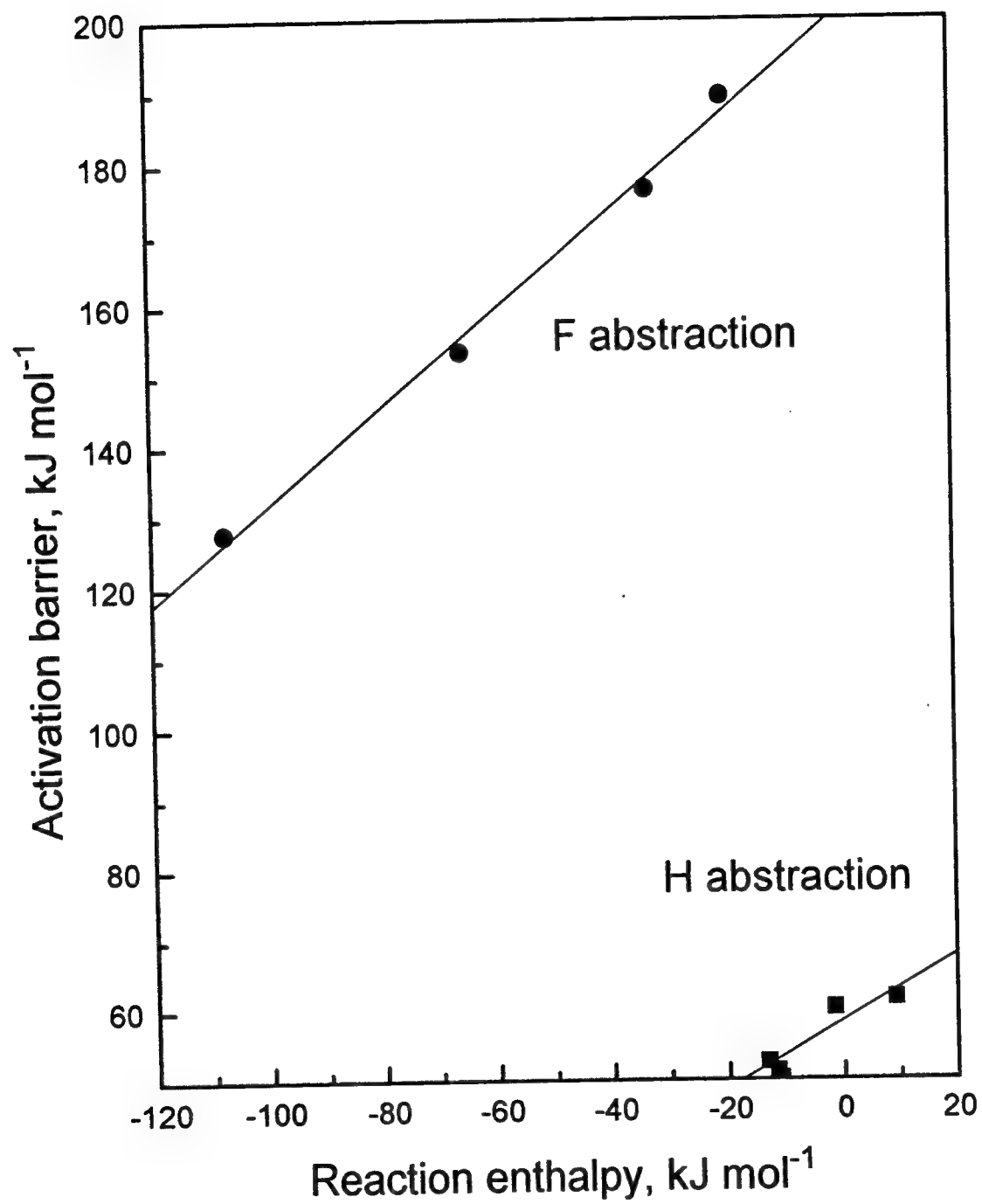


Fig. 3. Comparison between TST (heavy line) and experiment (ref. 18) for  $\text{H} + \text{CH}_4$ .

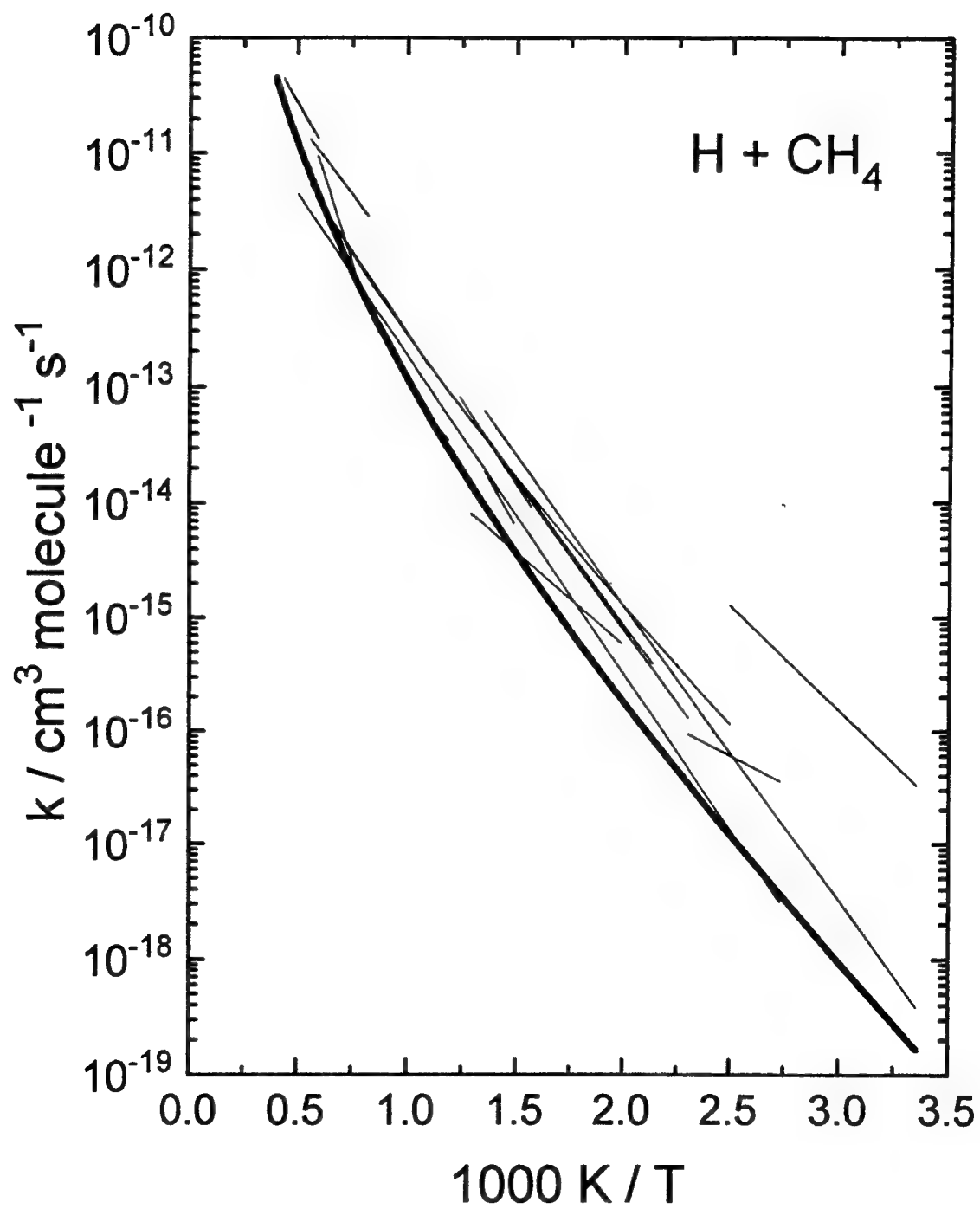


Fig. 4. Comparison between TST (heavy lines) and experiment (ref. 18) for  $\text{H} + \text{CH}_3\text{F}$ .

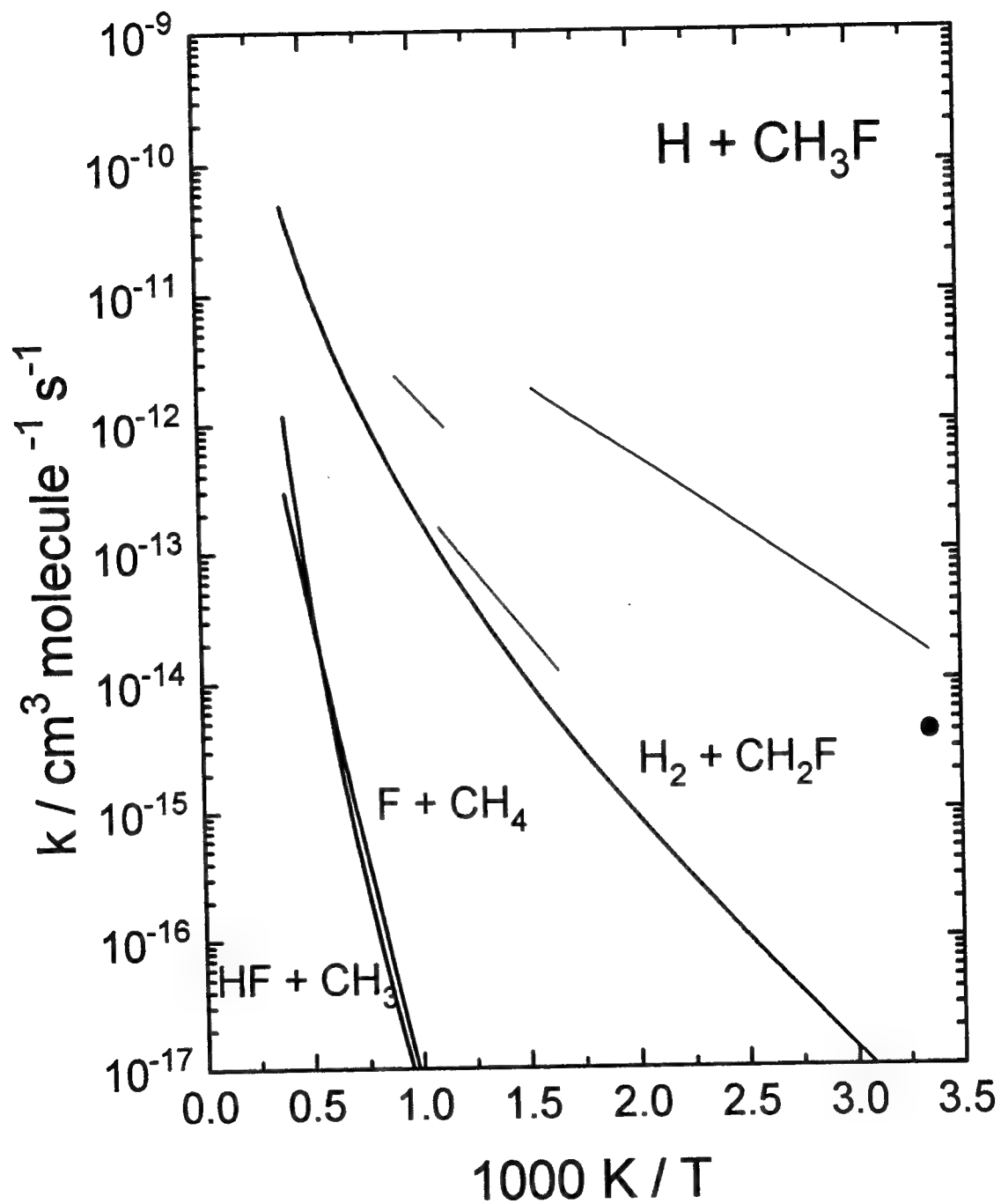


Fig. 5. TST results for  $\text{H} + \text{CH}_2\text{F}_2$ .

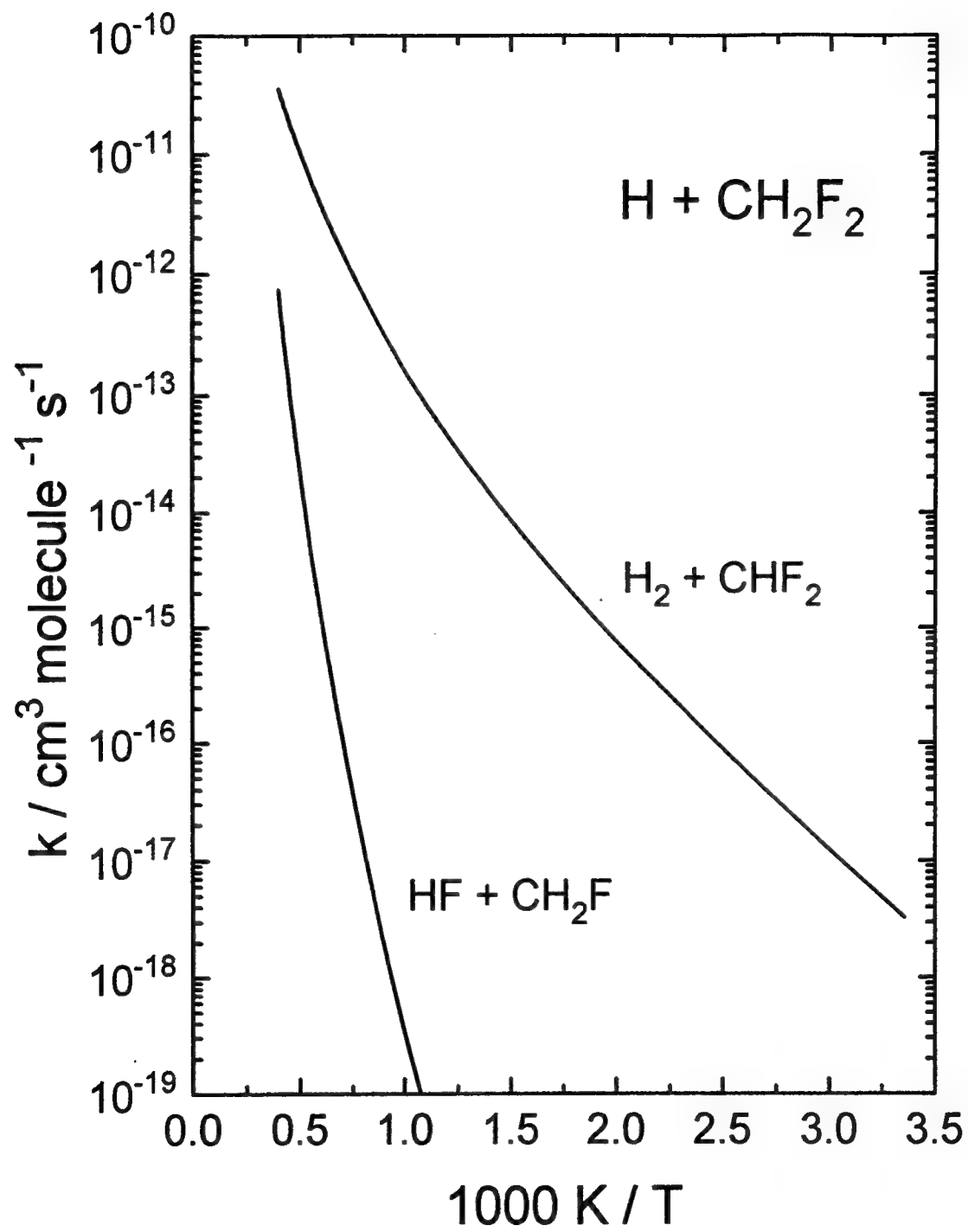


Fig. 6. Comparison between TST (heavy lines) and experiment (ref. 18) for  $\text{H} + \text{CHF}_3$ .

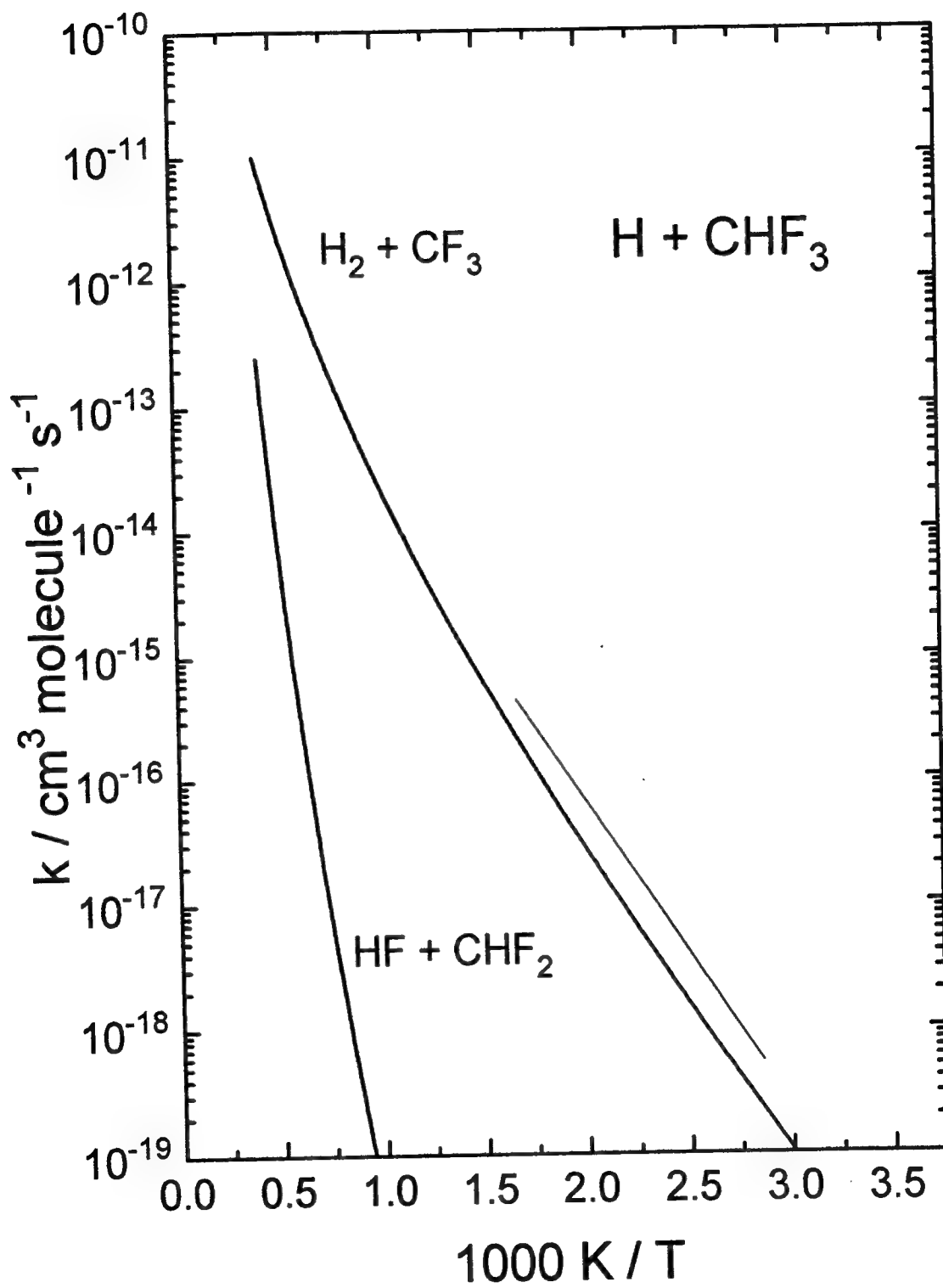
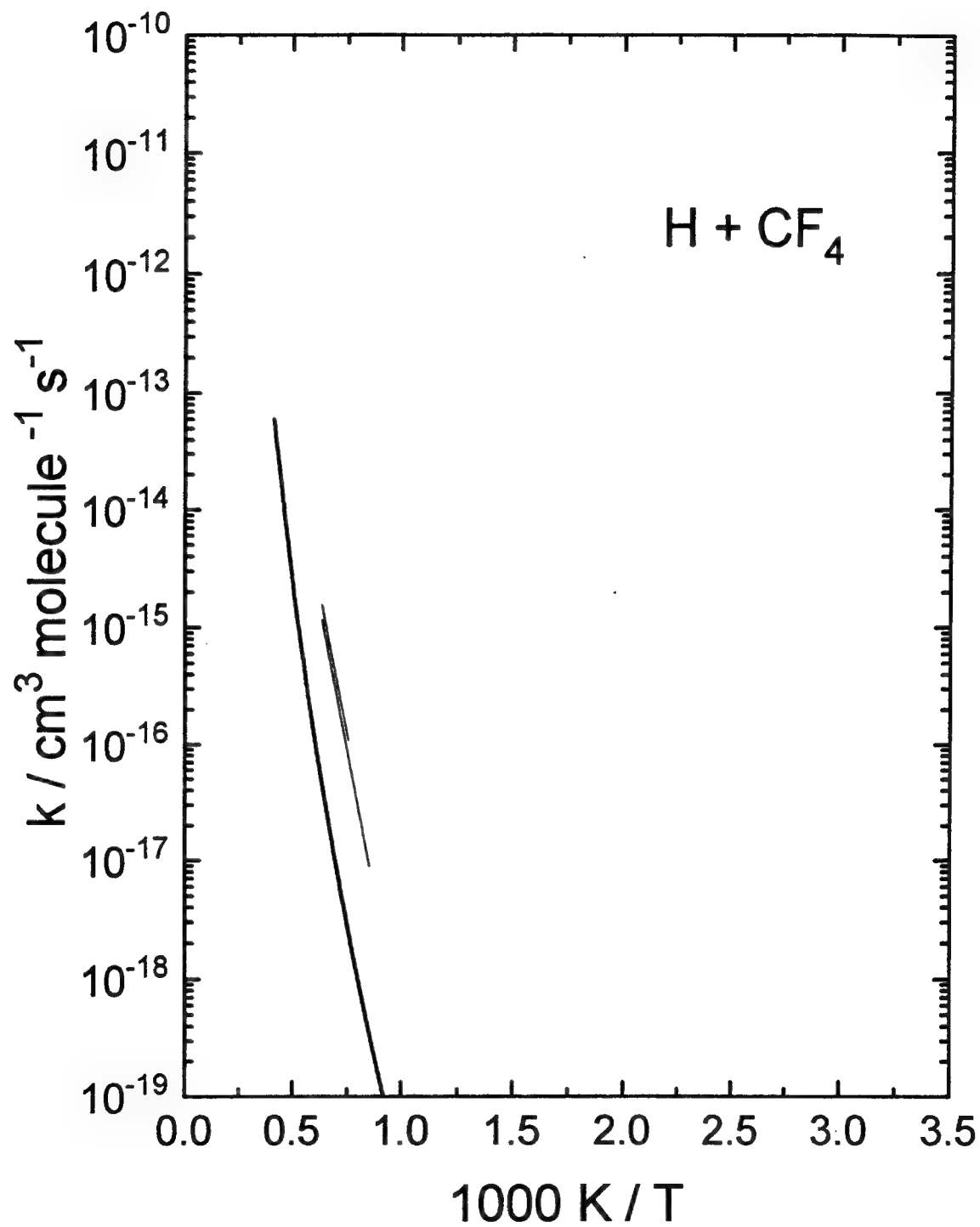




Fig. 7. Comparison between TST (heavy line) and experiment (ref. 18) for  $\text{H} + \text{CF}_4$ .



---

Preliminary Investigations of Mechanically Initiated Reactions  
in Energetic Materials

James J. Mason  
Clark Equipment Assistant Professor  
Department of Aerospace and Mechanical Engineering

University of Notre Dame  
365 Fitzpatrick Hall  
Notre Dame, IN 46556

Final Report for:  
Summer Faculty Research Program  
Wright Laboratory, Armament Directorate

Sponsored by:  
Air Force Office of Scientific Research  
Bolling Air Force Base, DC  
and  
Wright Laboratory

August, 1995

# PRELIMINARY INVESTIGATIONS OF MECHANICALLY INITIATED REACTIONS IN ENERGETIC MATERIALS

**James J. Mason**

**Clark Equipment Assistant Professor**  
Department of Aerospace and Mechanical Engineering  
University of Notre Dame  
Notre Dame, IN 46556

## **Abstract**

This report summarizes the activities of the author while studying the effects of dynamic mechanical loading upon ignition and reaction processes in energetic materials at the Advanced Warhead Evaluation Facility (AWEF) at Eglin A.F.B under the Summer Faculty Research Program in the summer of 1995. Some research at the AWEF is currently focused on the effects of dynamic stress state and material microstructure upon ignition and combustion processes in solids. Namely, the effects of dynamic compression, tension and shear loading, and material microstructure are being addressed with particular attention paid to the formation of thermal localizations during deformation. The shock to detonation transition (SDT) is not presently emphasized; rather, lower level stress states are of interest. The primary experimental research methods currently include the torsional split Hopkinson bar, the split Hopkinson pressure bar and the time resolved Taylor impact test. At AWEF experimental investigations are to be complemented by numerical calculations using finite element and finite difference methods. The results of the current work will be useful in understanding the interaction of material microstructure and mechanical properties with the kinetics of deflagration, detonation and initiation of energetic materials. Understanding of the mechanical and microstructural effects can lead to better safety, reliability and reproducibility of the performance of such materials in numerous devices and applications including deep earth and hard target penetrators.

# PRELIMINARY INVESTIGATIONS OF MECHANICALLY INITIATED REACTIONS IN ENERGETIC MATERIALS

James J. Mason

## 1. Introduction

By the estimation of the author, the purpose of the Faculty Summer Research Program, in terms of benefiting the faculty participants, is three-fold: to introduce a faculty member to the research interests of the Air Force, to make unique experimental capabilities available to the faculty member and to help the faculty member develop graduate student researchers at the same time through cooperative use of the Graduate Student Summer Research Program. Each of these objectives was addressed in the summer of 1995 while the author was in residence in Wright Laboratory, Armament Directorate, Munitions Division at Eglin AFB, FL. Most of the time in the summer was spent at the Advanced Warhead Evaluation Facility (AWEF) where discussions with Dr. J. Foster and Mr. D. Wagnon were made easier by proximity, but a familiarity with the interests and capability of other sections within the division was established through visits and ongoing discussions.

## 2. Research Interests of the Munitions Division

In the Munitions Division there is currently considerable interest in the mechanical initiation of explosive materials, and over the course of the summer a new understanding of initiation of solid explosives was gained by the author. Typically, a reaction in high explosives can be initiated by one of several mechanisms; thermal initiation, shock initiation and failure initiation. Each of these mechanisms is discussed below.

In addition to interest in mechanical initiation of explosives, there is also, in the Munitions Division, interest in dynamic fracture, dynamic failure of cylinders and dynamic crack arrest. Some understanding of the motivation behind this interest was gained by the author as discussed at the end of this section.

### 2.1 Thermal Initiation of Explosives

Generally, all initiation is considered thermal. For example, mechanical deformation leads to a temperature rise in the material, and that temperature rise causes initiation. Thermal initiation of explosives is generally the result of any temperature rise in a reactive material. It is frequently postulated that the important reaction occurs according to an Arrhenius relationship,

$$\dot{\lambda} = (1 - \lambda)\dot{\lambda}_0 e^{-\frac{E}{RT}}, \quad (1)$$

where  $\lambda$  is a reaction parameter giving the degree of completion of the reaction;  $\lambda = 0$  indicates no reaction has occurred and  $\lambda = 1$  indicates that the reaction has proceeded to completion. In this model the reaction is proceeding even at low temperature, but it does so at an infinitesimally small rate. As the temperature rises, the exothermic reaction proceeds at a faster rate until it reaches an instability or extremely high rate of reaction. The reaction then quickly proceeds through the material and may transition into a detonation. The transition is not always clearly understood, but it is considered to depend upon complex coupling of chemical

and mechanical effects. Modified Arrhenius equations are sometimes used to introduce pressure dependency in the reaction rate equation. Motivation for the introduction of pressure dependency in the reaction rate is born out by strand burner investigations of burn rates in solids.<sup>1</sup> Unfortunately, although there are many experimental investigations of burn rates using strand burner methods and other methods such as the closed bomb method, few other experimental investigations of burn rates in solids under shock and detonation conditions exist. The wedge test is the most commonly used method to examine detonation front advance under shock and detonation conditions.<sup>2</sup> Arrhenius kinetics are generally accepted for slow burning processes such as a burning strand, but their applicability in shock/detonation conditions is questionable; the validity of such equations is not fully characterized under those conditions.

The Forrest Fire model<sup>3</sup> removes direct temperature dependence in the reaction rate equation substituting pressure dependence instead. The unknown constants of the model are found from detonation wedge tests making it more suitable for shock and detonation conditions. But it is, at best, empirical. The Lee-Tarver equation<sup>4</sup> goes another step beyond the Forrest Fire model in that dependency upon specific volume is added. This model is intended to introduce hot spot initiation and growth into the reaction kinetics in a numerically simple manner. Many variations of the Lee-Tarver equation exist. It is difficult, however, to obtain the unknown materials constants for such models quickly and easily.

All reactions can be viewed as thermally initiated reactions, even those initiated mechanically. However, the study of reactive materials is essentially the interdisciplinary study of a chemically, thermomechanically coupled system. While the Forrest Fire model and Lee-Tarver type equations do not always include temperature dependency in their modeling of reaction rates, temperature is thermomechanically linked to pressure, and, therefore, its use may be justified in terms of thermal initiation in some situations. The difficulty in modeling the mechanical initiation of energetic materials lies in the complex link between chemistry and mechanics, especially solid mechanics.<sup>5</sup> Significant advances in the understanding the interaction of chemistry and mechanics for gaseous or liquid materials have been made, but much is left to be done regarding solid materials which, as daily experience demonstrates, behave quite differently. In light of the complexity involved, thermal initiation is often simplistically described by a critical temperature criterion. If the temperature exceeds some characteristic critical temperature,  $T_c$ , then the reaction will initiate.<sup>6</sup>

## 2.2 Shock Initiation of Explosives

Shock initiation of gases and liquids has received considerable attention over the last century. This type of initiation results from the rapid application of extremely high pressures,  $\geq 100\text{kbar} = 10\text{GPa}$ , or from impact at very high velocities,  $\geq 1\text{km/s}$ . The mechanism is well described by the conservation equations of continuum mechanics; the conservation of mass,

$$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho v_j)}{\partial x_j} = 0, \quad (2)$$

the conservation of energy,

$$\rho \frac{DE}{Dt} = k \nabla^2 T + p \frac{\partial V}{\partial t} + q, \quad (3)$$

and the conservation of momentum,

$$\rho \frac{Dv_j}{Dt} = \rho b_j - \frac{\partial p}{\partial x_j}. \quad (4)$$

These equations are referred to as the continuity equation, the energy equation and the equation of motion, respectively. The density is  $\rho$ ,  $v_j$  is the velocity component,  $E$  is the internal energy,  $k$  is the thermal conductivity,  $T$  is temperature,  $p$  is the pressure,  $V$  is the specific volume,  $q$  is the heat of reaction per unit mass, and  $b_j$  is a body force. The material is assumed to be a non-viscous fluid that obeys Fourier heat conduction. Then, it is assumed that a shock forms in the material and reactions occur due to compressive heating of the material behind the shock, a result of thermomechanical coupling in the equations. In one simple model the reaction rates are assumed infinite, the conservation equations are applied at the shock front, and the equations for conservation of mass and momentum are combined and represented as the *Rayleigh Line* in  $p$ - $v$  space,<sup>7</sup>

$$p_2 - p_1 = -\rho_1^2 v_s^2 (V_2 - V_1)$$

where the 1 and 2 refer to material ahead of the shock front and material behind the front, respectively. The shock velocity is  $v_s$ . The conservation of energy, mass and momentum across the shock front may be stated in the form of the *Hugoniot Curve* in  $p$ - $v$  space

$$e_2 - e_1 = \frac{1}{2}(p_2 + p_1)(V_1 - V_2) + q.$$

These are two equations for four unknown variables,  $e_2$ ,  $p_2$ ,  $V_2$  and  $v_s$ . With a caloric equation of state for the material behind the front, the internal energy,  $e_2$ , may be expressed in terms of  $(p_2, V_2)$  reducing the number of unknown variables in the Hugoniot and Rayleigh equations to three.\* The *Rayleigh Line* and *Hugoniot Curve* are plotted in  $p$ - $V$  space and solutions dependent upon  $v_s$  are found where the two curves intersect. This approach is generally referred to as Chapman-Jouguet (CJ) Theory, and when the Rayleigh Line is tangent to the Hugoniot the intersection is called a CJ point. Two such points exist, called the upper and lower CJ points as shown in Figure 1. The upper point gives the detonation velocity for the shocked material, the minimum velocity at which a detonation can propagate and satisfy the given equations. At this point it is understood that shock compression heats the material to induce an instantaneous reaction. CJ theory leaves much of the reaction and deformation process unexamined, and more complicated analyses are required to handle the addition of finite rate reaction kinetics, and the inclusion of viscosity which requires a new constitutive law and more complicated thermal equation of state.<sup>9</sup> However, it is simple, and, consequently, in what follows the discussion will be limited CJ theory and to how it relates to lower stress level initiation of solids.

The system of equations in CJ theory exhibits thermomechanical coupling, the conversion of mechanical energy to thermal energy and vice versa, in the equation of energy and the equation of state in a way that is typical for compressible fluid mechanics formulations, through volumetric changes in the material. Solution of these equations would be greatly simplified if the material were assumed to be incompressible,<sup>10</sup> then  $\rho$  is constant and  $\partial v / \partial t = 0$ . The energy equation would then be uncoupled from the equation of motion, and

---

\* Equations of state typically come in two forms; the thermal equation of state, which is essentially a constitutive law, and the caloric equation of state, which relates the internal energy of the material to the unknown variables of the problem. For a perfect gas one might assume the following thermal equation of state and caloric equation of state respectively;<sup>8</sup>

$$p = \frac{RT}{V},$$

$$e = c_v T.$$

The second equation, or the caloric equation of state, may be stated in terms of  $p$  and  $V$  by substituting the thermal equation of state in the caloric equation. Variations of the perfect gas law are often used.

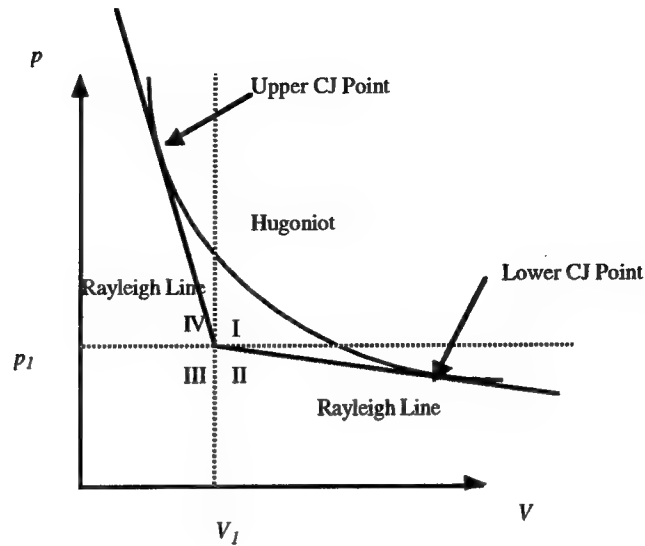


FIGURE 1 Diagram of the Rayleigh Lines and Hugoniot Curve.

solutions could be found without invoking thermomechanics. However, it is necessary to include the coupling in order to accurately explain observed reaction behavior. In fact, this approach is fundamentally related to the field of compressible fluid mechanics and gas dynamics. Note that none of the familiar constitutive laws from solid mechanics—elasticity, viscoelasticity, plasticity, viscoplasticity, etc.—are needed. In the formulation of this model no provision was made for material strength, but because the pressure and stress due to shocks far exceed the strength of solids, the strength is negligible and the model may be used for such solids. Physically, this is understandable because the impact velocities and/or resultant pressures far exceed the elastic range of behavior of the material and far exceed the yield strength of the material so that the deformation proceeds along micromechanical mechanisms that are more similar to mechanisms of deformation in a fluid than in a solid. Although the material may have an initial crystalline structure, that structure has no dominant effect on the mechanical response. Only the current specific volume and pressure are required to characterize the material stress state and deformation state. Mathematically, the lack of a stress tensor and related constitutive equations simplifies the problem greatly. Stress becomes an isotropic tensor described by the scalar pressure (disregarding viscosity effects, of course), reducing the number of unknown stress components to one. Deformation is described by the specific volume, a scalar, which reduces the number of unknown deformation measures to one.

It is often assumed that plasticity in solids is incompressible, a constant volume process. This might lead one to believe from CJ theory and similar fluid mechanics based theories that plasticity doesn't enter into solid detonation phenomena. The equations become thermomechanically uncoupled for incompressible materials. But, at the level of loading described by shock initiation, the familiar deformation mechanisms associated with plasticity are not dominant. The material is most likely compressible, and, although it probably is not well modeled by the perfect gas law, it has been observed to behave more like a fluid. So, invoking plasticity in shock phenomena is totally inappropriate. Only when the load levels are decreased does incompressible plasticity enter the problem. Then, the thermomechanical coupling takes a different form, strength is also significant, and an entirely new problem formulation is required.

Furthermore, when plasticity is possible many other types of material behavior that lead to mechanical

failure of the solid are also possible. For that reason lower stress level ignition phenomena in solid explosives are called "failure initiation" mechanisms by the author.

### 2.3 Failure Initiation of Explosives

Failure initiation, in contrast to shock initiation, can be more complicated mathematically and includes aspects of both shock initiation and thermal initiation. The material is mechanically loaded, as in shock initiation, but at lower levels,  $\approx 100 \text{ Mpa} = 1 \text{ kbar}$ , which indicates that the usual constitutive laws of solid mechanics apply. There may be plasticity, elasticity, fracture, fragmentation, stress concentrators, melting, phase transformation and unloading (which is different than loading in many solids), and the stress becomes a tensor with six independent components. Also, the deformation is no longer well described by specific volume alone; the thermodynamic counter part to stress, strain or strain rate, must be introduced as a tensor adding five more unknowns. The new constitutive behavior increases the number of unknowns by ten. Using tensor notation the governing conservation laws can be written as; continuity,

$$0 = \frac{\partial \rho}{\partial t} + \frac{\partial(\rho v_j^p)}{\partial x_j}, \quad (5)$$

energy,

$$\rho \frac{De}{Dt} = \sigma_{ij} v_{i,j}^p + k \nabla^2 T + q, \quad (6)$$

and momentum,

$$\rho \frac{Dv_i}{Dt} = \sigma_{ij,j} + \rho b_i, \quad (7)$$

where now there are 5 equations and 12 unknown variables.\*\* The plastic deformation rate tensor,  $v_{i,j}^p$ , has been introduced such that  $v_{i,i}^p = 0$  indicating incompressible plasticity. For the sake of simplicity, elasticity has been neglected. The constitutive law (stated in terms of velocities) will give 6 more equations, plus a caloric equation of state will give the seventh making the system well defined. Obviously, the constitutive laws, which make up more than half the equations in the system, dominate the response of the system.† The material behavior can be quite complicated i.e. leading to shear localization, fracture, void nucleation and growth, fragmentation and friction between failed surfaces. Typically, except for simple loading geometry and material behavior, this system of equations is solved numerically using finite element or finite difference schemes—as is done for the case of viscous fluids.

Quite often in solid mechanics there is no mention of the caloric equation of state. This equation may be neglected if the energy equation is not included, or it may be directly substituted into the energy equation. Often the energy equation is neglected in solid mechanics since isothermal conditions are assumed—not realistic for high rate loading—or thermomechanical coupling is deemed negligible—despite experimental evidence to the contrary. In the energy equation there is thermomechanical coupling similar to that in shock initiation in the previous section, but the thermomechanical coupling appears in the form of plastic heating due to shear,  $\sigma_{ij} v_{i,j}^p$ , rather than volumetric work,  $p \frac{\partial v}{\partial t}$ . Plasticity is frequently approximated as incompressible which simplifies the equation for conservation of mass ( $\rho = \text{const.}$ ), but for void filled materials this may not be an accurate approximation.

\*\* No shock has been assumed to form so comparison to the number of equations and unknown variables in CJ theory is not valid.

† Sometimes the constitutive law must be stated in terms of displacements and velocities which adds even more equations and unknown variables!



The initiation mechanism is most simply connected to thermal initiation through the use of a critical initiation temperature criterion.<sup>6</sup> The system of equations are solved and when the critical temperature is reached, initiation is begun. But, for sophisticated failure initiation criteria, the reaction kinetics will be required. Certainly, the initiation criteria must include temperature and time of exposure,<sup>6</sup> but, if numerical simulations are to be used, reaction kinetics may be modeled numerically using the thermal initiation/reaction models described in Section 2.1.

At AWEF the principle interest in failure initiation of explosives is motivated by safety considerations in the storage of such materials and the engineering design of hard target and deep earth penetrators. Hard target and deep earth penetrators are to be designed to penetrate heavily enforced targets such as concrete bunkers buried deep underground and other heavily reinforced buildings. These penetrators will be required to undergo severe mechanical insults before they reach the target and explode. Concerns about premature initiation of the material are well founded. Also, if the explosive material is fragmented upon impact before reaching the target, it may not perform properly upon arrival, and concerns about proper detonation at the target are also well founded.

#### *2.4 Systems and Mechanisms of Interest*

The materials of interest to the Munitions directorate fall into two categories, granular or polymeric. The granular materials, such as Nitonal or 9501, are characterized by low moduli, low strength and low deformation to failure when loaded under low hydrostatic stress. However, it is thought that when the hydrostatic stress, i.e. pressure is high, the strain to failure can increase dramatically and the materials may be considered ductile rather than brittle.<sup>11</sup> These materials resemble geological materials in mechanical behavior and microstructure. The polymeric materials, plastically bonded explosives (PBX's), such as PBXN9 and PBXN109 are characterized by low moduli, low strength, possibly viscoelastic behavior, and large strain to failure. These materials resemble medium to hard rubber.

One of the problems involved with performing numerical simulations of the failure initiation of these materials is the lack of accurate constitutive laws. Few complete characterizations of the materials mentioned above are available. Characterization requires an extensive testing procedure using a variety of techniques and apparatuses in order to get a meaningful constitutive law that is valid for any complex loading geometry over a wide range of strain rate, temperature and strain. This is needed to accurately describe the deformation of the materials of interest under all probable loading conditions. If failure other than plastic yielding is of interest further investigations into the failure mechanism under a host of conditions is required. The task is very large indeed. Techniques such as torsional Hopkinson bar, split Hopkinson compression/tension bar and Taylor impact testing may be used to characterize the materials at high strain rate.<sup>12</sup> Currently, testing related to the project underway is being performed using the time resolved Taylor impact test at AWEF, using a split Hopkinson pressure bars at Los Alamos National Lab. and using a punch test<sup>11</sup> and a modified Taylor impact test<sup>13</sup> at Dyna East. In addition, testing at low strain rate using servo-hydraulic testing machines is underway. However, none of these techniques involves high hydrostatic pressures, except perhaps the modified Taylor impact tests.

Concerns over failure initiation upon impact lead to an interest in several possible initiation mechanisms. These mechanisms are usually limited to deformation or temperature localization mechanisms because localization can lead to ignition when homogeneous deformation will not. Consequently, hot spots are seen to be the most probable initiation sights. The possible mechanisms for the formation of a localization or hot spot include shear localization, deformation concentrators such as voids or hard particles and fracture

or fragmentation (which leads to frictional heating).

(a) *Shear Localization*

Shear localization is basically an instability in the system of equations that occurs when thermal softening of the material (strength decreases with temperature in most materials) and thermomechanical coupling dominate the deformation. This is a result of constitutive behavior. Under conditions of monotonically increasing deformation a decrease in strength leads to further deformation producing heat that induces more deformation and so forth. Traditionally, solid mechanics has assumed isothermal conditions or a lack of thermomechanical coupling in its models. Under such conditions shear localization is not possible. However, it can be shown that the introduction of thermomechanical coupling leads to such instabilities under certain conditions. For example Clifton et al.<sup>14</sup> showed the possibility for instability by starting with a mathematically stable homogeneous deformation for an uncoupled system and showing that such a deformation was unstable under certain conditions after coupling was introduced. But, in the coupled system the homogeneous deformation is not a good starting point under fixed temperature boundary conditions, it simply does not satisfy the equations. The initial condition is therefore inherently unstable, and must evolve to a steady state configuration. It remains to be determined what the steady state solutions are. And also, how the system evolves to those steady state solutions. Ideally there will be one solution that reflects shear localization and one steady state solution that doesn't. Further work to find the nature of these solutions is necessary.

Shear localization is of interest because it can lead to very high temperatures which then produce initiation. In metals Marchand and Duffy<sup>15</sup> have reported a lower bound on the temperature in a shear band to be approximately 600°C. It has also been observed that martensitic transformations occur within shear bands in steels<sup>16</sup> indicating that the temperatures may be much higher. Chou<sup>11</sup> indicated that shear localization and initiation occurred in computational simulations of the impact of explosive materials leading to temperatures that could initiate a reaction in the material, but this only occurred when the material was constrained by armor. Regardless, shear localization is a precursor to fracture, and fracture ends the shear localization event thereby ending its associated heating. Therefore, fracture should be included in the model because it may have a significant effect upon the initiation event. However, up to this point in time, fracture has not been included in numerical simulations of shear localization. Perhaps because fracture is difficult to model or because it may also generate heat. Zehnder and Rosakis<sup>17</sup> and Mason and Rosakis<sup>18</sup> observed temperatures around 500 °C and 300 °C at the tip of dynamic crack tips in steel and titanium, respectively. These temperatures may be sufficient to initiate a reaction as well. In summary, it is important to know how soon fracture follows shear localization and what effects the fracture event has on the heating of the material in order to correctly predict initiation of reactions by shear localization. For that reason some work was performed by the author to examine the failure within a shear localization. That work is discussed in the following section on experimental investigations.

Variations of microstructure can enhance shear localization. Zhou et al.<sup>19,20</sup> have shown that in tungsten heavy alloy materials inhomogeneity in the material properties leads to earlier shear localization. The tungsten heavy alloy is essentially a composite material with matrix and reinforcement. These authors tested the composite material and the matrix/reinforcement materials in a plate impact geometry. It was observed that shear localization occurred earlier in the composite than in the matrix or reinforcement when tested separately. This is attributed to the microstructural variation in material properties within the composite. The variation of strength, thermal conductivity, elastic moduli, etc., contributes a perturbation in the

deformation field which essentially leads away from an unstable state, quasi-homogeneous deformation, to a stable state, shear localization. The perturbation in microstructure is strong and unstable and, therefore, leads to localization earlier in the deformation. There may be, however, a microstructure that stabilized the homogeneous state leading to a shear-localization-resistant composite. The conclusion to be drawn from the work of Zhou et al. is that effects of microstructure upon localization can be significant and careful attention should be paid to the microstructure.

*(b) Stress/Deformation Concentration*

Deformation concentrators such as hard particles and voids have been known to sensitize reactive materials for many years.<sup>6</sup> However, the exact mechanism and deformation leading to initiation is not always clear. Deformation may be concentrated by the geometry of the loading or the geometry of the problem. The stress concentration leads to ignition by one of many mechanisms including: void collapse, fracture, shear localization, friction between grains or friction between inclusions and matrix materials. It has been postulated that intense plastic deformation around a particle leads to initiation or that void collapse leads to adiabatic heating of the gas/material within the void or that collapse of the void leads to shock initiation of the material due to impact of opposing faces or that friction between fractured surfaces leads to ignition or that shear localizations appear between stress concentrators or that plastic deformation at a generated crack tip leads to ignition and so on. Certainly, materials characterization combined with numerical simulation and experimental investigation would clarify the importance of each of these mechanisms for the target materials.

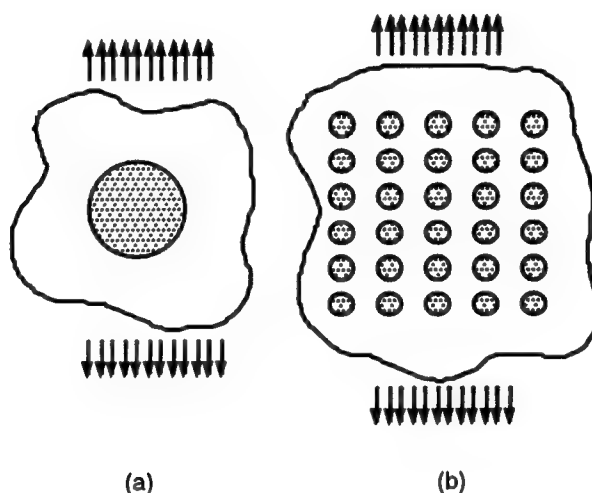


FIGURE 2 Schematic of two analytical problems that could be addressed using numerical techniques.

There has been a large amount of analytical modeling, whether numerical or closed form, of void collapse and its effect upon energetic material ignition.<sup>2,6,21-25</sup> However, surprisingly little has been done in regard to the micromechanical effects of the inclusion of hard particles in an energetic material or the micromechanical effects of mixing two phase energetic materials of different mechanical properties.<sup>†</sup> Analytical efforts should focus on the micromechanics of hard particle inclusions and two phase mixtures in mechanical ignition of

<sup>†</sup> There is a good amount of work done on the bulk reactive and detonation properties of two phase energetic materials, however.<sup>26-28</sup>

solids. Two problems could be addressed as shown schematically in Figure 2. First, a single spherical inclusion, as shown in Figure 2(a), could be examined under conditions of dynamic shear, compression and tension. Of particular interest in this problem is the role of the interface strength, the particle properties, thermal softening in either material, friction between the particle and matrix and the fracture properties of both materials. It is of interest to examine both the deformation of the material surrounding the particle and of the particle itself under the different loading conditions. The interfacial strength will play an important role since debonding can lead to a host of different effects. For example, friction between the particle and matrix after debonding may be the dominant mode of heat generation. Some effort could also be made to examine the role of particle morphology. Particularly, the existence of sharp corners in the particles will lead to stress concentration and it would be interesting to quantify this effect. A hexagonal particle as shown in Figure 3 could be used under the same conditions of loading as earlier investigations to quantify this effect.

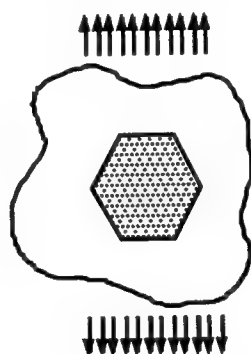


FIGURE 3 The effects of sharp corners such as those shown here should be addressed.

Next, an array of included particles, as shown in Figure 2(b), could be examined under conditions of dynamic shear, compression and tension in an attempt to produce a more macroscopic, maybe asymptotic, characterization of the behavior of a particle array. The results could be used in a large scale numerical scheme to represent the composite behavior of the material in a computationally efficient manner. In this formulation a regular array could be examined to see the effects of particle spacing on deformation localization. Then, a random array could be examined to demonstrate the effects of uneven particle spacing. As in the first case, particular interest should be direct toward the role of the interface strength, the particle properties, thermal softening in either material, friction between the particle and matrix and the fracture properties of both materials. However, the interaction between particles should be emphasized and the limits of the particle density should be explored. As particle density gets high analogies between the numerical model and two phase bulk materials can be made.

The results of the numerical work should be coupled to experimental work where appropriate. It is hoped that numerical investigations of the micromechanisms in general will result in intuitive understanding of experimental results. Then, with some additional materials characterization, the numerical codes may be used to predict, or at least reflect, the observed behavior giving valuable insight into the relationship between microstructure and material response.

### (c) *Fracture Related Events*

Concerns about maintaining the integrity of the high explosive, so that it will properly detonate when it reaches the target, lead to interest in fracture and fragmentation of the material. Creation of a damaged

bed can lead to a completely different combustion behavior when compared to the solid. But, fragmentation and fracture are difficult to model numerically, and frictional effects in the bed may be important. Modeling of this type of behavior may not be possible using continuum mechanics equations, Equations (5)–(7), unless some sweeping assumptions can be made about the frictional, mechanical and kinematic behavior of the particles. The fragmentation event itself is also difficult to model, and current numerical methods typically use simple element elimination routines to mimic fragmentation behavior.<sup>29</sup> It would be worthwhile to develop a more sophisticated numerical scheme that can be implemented in standard finite element and finite difference codes.

*(d) Summary of Failure Ignition Mechanisms*

In what preceded only three failure ignition mechanisms have been identified and discussed; shear localization, stress concentrators and fragmentation/fracture. Many others may exist, and, because there are so many possible ignition scenarios, there is a need to identify the most important/dominant mechanism for each study material. Various criteria for ignition under the different mechanisms are required so that the mechanisms can be compared subjectively to determine which might become active. Considerable effort will be required to produce such criteria; in fact, the feasibility of such an undertaking might be questionable. Some critical experiments will certainly be required.

*2.5 Dynamic Fracture of Cylinders*

In another section of the Munitions division, the High Explosive Research and Development section (HERD), considerable interest has been generated in the dynamic fracture of internally pressurized tubes. Because most explosive devices or bombs are essentially dynamically pressurized cylinders, it stands to reason that the fracture behavior of such cylinders will greatly effect the performance of such devices. In particular, it is sometimes desirable to explosively cut a device in half. Then the aft section of the device may be detonated while the fore section continues on its path to another target. At the HERD just such a device is being developed, however problems due to axial crack propagation have arisen. The situation is schematically illustrated in Figure 4. While the objective is to fail the cylinder circumferentially, the loading more readily lends itself to crack propagation axially, as shown. There are two strategies to avoid axial growth; one is to change the material properties so that axial crack initiation is delayed until circumferential failure is imminent and the other is to allow axial crack to nucleate but then arrest their growth. Changing the materials properties is not uncommon, but in terms of changing the properties to retard crack nucleation, this strategy may not be fruitful. Retarding axial crack nucleation may retard circumferential failure as well, and inhibiting crack nucleation means using more ductile, lower strength, forged, clean materials. Not only is this expensive, but it also contradicts other design criteria. For example, a high strength material is desired to survive target penetration while a low strength steel is desired to inhibit circumferential failure. In addition, if materials were to be changed or a different device were to be used, more lengthy alloy development would be necessary. For that reason it seems more logical to attempt to arrest the cracks after they nucleate. This may be achieved by changing the geometry of the device, not by changing the material.

The pipeline industry has been faced with the problem of arresting dynamically propagating axial cracks in an internally pressurized cylinder for many years. Typically, this industry will attach a crack arrestor to the pipeline at some interval along the length of the pipe. Without such arrestors an axially propagating crack may run for miles in the pipeline. The reason for such propagation is rather simple. The internal pressurization is usually a gas or liquid. (In the case of an explosive device it is probably mostly gaseous

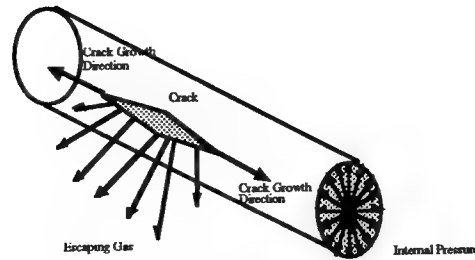


FIGURE 4 An axial crack extending in an internally pressurized cylinder.

reaction products.) If a crack is nucleated it will propagate at a speed of roughly 1000 m/s in steel. Because the internal pressurization stores elastic energy in the cylinder the crack may propagate indefinitely by using the stored energy to propagate. The only way the system, as described, can release the stored energy before the crack can use it is to release pressure. This may occur by releasing the gas through the open crack. However, the unloading wave, due to the escape of the gas, travels at the speed of sound in the gas, roughly 300 m/s in air. Thus, the effects of the unloading due to venting of gas cannot reach the 1000 m/s crack tip. The crack tip outruns the unloading mechanism and propagates into a fully pressurized cylinder until some geometry change prohibits its propagation. There is a need to develop a crack arrestor for the particular geometry being studied at the HERD.

### 3. Use of Experimental Capabilities by the Author

Three experimental investigations were started at AWEF. One was carried out by the author and the remaining two were addressed by graduate students.

The author became interested in fracture events that follow shear localization because of the implications this event can have on initiation of solid explosives. In metals Duffy<sup>15</sup> and Giovanola<sup>30</sup> observed the onset of failure within shear localization bands in torsional Hopkinson bar tests and both authors report a three stage process; shear localization, collapse and fracture. But neither author characterized the growth of the crack that followed shear localization other than indicating that the crack growth initiated randomly within the shear band somewhere around the circumference of the cylinder. Since no provision was made to observe the entire cylinder during deformation and failure, the actual crack nucleation event was not observed. Any attempt to observe it would be hindered by the random nature of the nucleation. A crack could nucleate anywhere on the circumference of the cylinder depending upon the wall thickness variation, temperature variation or microstructural variation in the material. The location of the nucleation site is very difficult to predict. What is needed is a geometry in which the location of the nucleation event can be predicted with more certainty.

Shear localization at notch tips has been observed by Kalthoff,<sup>31</sup> Kalthoff and Winkler,<sup>32</sup> and Mason et al.<sup>33</sup> for the geometry shown in Figure 5. In this geometry a crack following shear localization has been observed to consistently initiate at the tip of a notch. The notch/crack is loaded dynamically in shear (mode

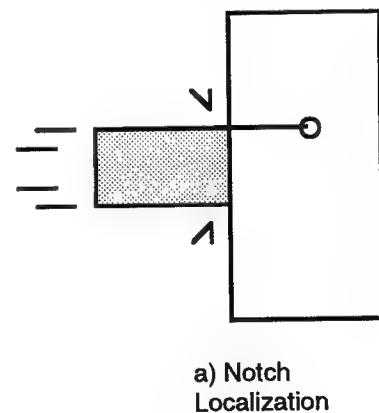


FIGURE 5

II) in the first few microseconds after impact. However, after some time the crack tip loading mode changes to mixed mode, mode I and mode II.<sup>34,35</sup> No direct observation of the crack initiation and growth process has been made. Although, after the test a crack is observed to have propagated along a path of intense shear and little if any shear deformation is seen ahead of the crack. It appears that the shear localization zone alone is completely failed, and the crack did not grow beyond the limits of the shear band. Sometimes, depending upon aging condition in the material or the projectile length, the crack changes path following a more tensile, rather than shear, failure mode.<sup>33,34</sup> First the crack grows straight ahead, then at later times it grows at an angle. This behavior is of interest because in one test it may be possible to discern the failure criterion for both shear crack advance and tensile crack advance.

Observation of the crack initiation and growth process in this geometry has not been reported in the past because most solid mechanics labs in university settings have only laser illumination for use in ultra-high-speed interferometry applications; photoelasticity, diffraction Moire and so forth, not straight photography. At the AWEF there is a white light source which allows full frontal illumination of the specimen with white light so that the crack initiation and growth can be recorded. But, there are some problems. First, the field of view needs to be small. Proper optics must be used to get magnification while maintaining a suitable stand off distance from the specimen. (After impact the specimen and plate fly into a momentum trap, however there is a chance that the optics could be damaged if they were positioned too close to the impact event.) Second, the suitable optics tend to have high f number. Light is therefore limited, and the white light source must be focused to concentrate the illumination on the specimen in the area of view. These two factors were addressed in the research performed at AWEF by the author. A Fresnel lens was used to focus the white light source. This procedure had been used in the past by technicians at the AWEF and, in fact, had also been suggested by the manufacturer of the high speed camera, Cordin. The camera is required to take photographs at one hundred thousand frames per second during one test to capture the later time behavior. Then in another test it is necessary to take pictures at six hundred thousand frames per second to capture the shear localization and failure. Two tests were to be performed. In the first test the change of crack propagation mode from shear to tensile growth mode which takes place over longer time scales, perhaps as much as 500  $\mu s$  (the time of the event was not known) was to be recorded. In the second test the initiation of the crack which takes place over very short time frame, 50  $\mu s$ , was to be recorded. Three optical systems were targeted for use in these experiments; a macro lens from Pentax, a micro lens from Nikon and a microscope objective with bellows from Nikon. After some experimentation, the bellows set up was abandoned in the interests of ease of use and time savings.

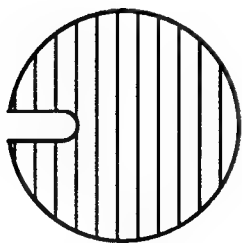


FIGURE 6

Specimens were prepared using water jet machining at the model shop on base at Eglin AFB. This technique was not satisfactory because the notch width varied through the thickness. A second set of specimens was sent out to an external machine shop to be prepared using wire EDM machining with much better results. The specimens were polished on a belt sander with the scratch marks running perpendicular to the expected shear crack propagation. This was done to make visualization of the shear deformation easier. In the past, photolithography has been used to deposit grids on shear specimens,<sup>15,30</sup> but in the interest of time that technique was not developed. Instead the scratch marks from polishing with coarse grit sand paper were used as fiducials. Projectiles were fashion out of 1/2" diameter 4340 steel rod.

Preliminary test were performed using the Cordin camera and focused light source on a stationary

specimen: no impact event was generated. This was performed to determine whether the light source was powerful enough to expose the film at the required framing rate and to determine if the grinding marks could be properly imaged on the film. The test was successful. The film was exposed at the highest required framing rate and the vertical marks were clearly imaged. Next, the impact event was to be generated. Unfortunately, the arrival of hurricane Erin in the area made the execution of this series of tests impossible. The lack of personnel due to various incurrences of property damage made it impossible to complete the tests at the originally scheduled time. It was necessary to reschedule time for the experiments at a later date, but the contracted tour of the author under the Summer Faculty Research program had been completed before such time could be arranged. Presently, the author is exploring the possibility of running the tests at Notre Dame or having the personal at AWEF run the tests in his absence.

#### 4. Related Graduate Student Research

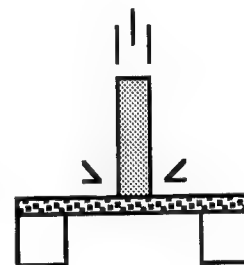
While in residence at the AWEF the author also supervised and advised two graduate students from the author's home institution who were working under the Graduate Student Summer Research program.

##### 4.1 Punch Tests

In work related to the investigations of the author and the interests of the Munitions Division, K. Roessig, under supervision of the author, performed plugging/punch tests on two materials, aluminum and steel. The purpose of these test was to begin understanding the punch test itself. The punch test, as shown in Figure 7, is nothing new, manufacturers have been punching metals for more than a century. However, it is surprising to note that little work has been done on high speed punching. Of the parameters that are important in punching three can be singled out as dominant; the speed of the punch, the radius of curvature of the punch (its sharpness) and the clearance between punch and die. A little work has been done on investigating the effects of high punch speed upon process and a large amount of work has been performed to investigate the effects of clearance and sharpness. But, little or no work has been done to investigate the interaction between clearance, sharpness and velocity. It is known that at very high punch velocity plugging will occur; that is, a nicely formed shear plug will be punched out of a plate. This occurs for effectively infinite clearance. It is also known that at quasistatic punch velocity a plug is nicely formed if the clearance is effectively zero. The transition from zero clearance to infinite in terms of the punch velocity required to get a nicely formed plug however has not been characterized. Roessig started to do that characterization at AWEF.

But, there were also other motivations for his work. The punch test is gaining favor as a means of testing the shear sensitivity of explosives. For example, see Chou<sup>11</sup> and Frey<sup>36</sup>. We expect that in the future a form of the punch test will be used to generate explosive initiation data under direct observation. And, this present work was seen as the start of an evolution in that direction.

Roessig built an entirely new apparatus, a die and supports, to be used with a 30 mm gun at AWEF. The gun is capable of muzzle velocities in the range from 100 m/s up to a few km/s. It was expected that the plugging velocity of the material would be in this range. So he set out to find the plugging velocity of



b) Punch Test

FIGURE 7



aluminum and steel. The apparatus was machined by the model shop and welded at AWEF. He encountered many difficulties during testing due in large part to the temporary lack of usable gun powder at AWEF and the difficulty of using the 30 mm gun to fire the punch at velocities near and below 100 m/s, its lower limit of performance. He was however able to fire the punch at velocities ranging above 60 m/s and find an upper bound on the plugging velocity, the velocity below which shear failure does not occur. More details of this work are reported in Roessig's Graduate Student Summer Research Report, base number 30. With the experience he has gained he will be well prepared to begin design of an apparatus, including a gun, to be used at Notre Dame in the completion of this project. He will build an air gun to launch the punch at 10 m/s to 300 m/s and finish this work at Notre Dame.

The punching and blanking tests were a success because we were able to carry out tests at velocities well beyond the current capacity at Notre Dame. In the process it was learned that the punching velocity is probably within the capabilities of the Notre Dame facilities for aluminum and steel and that an air gun, rather than a powder gun, is better suited for punch velocity tests in the range of 10 m/s to 1000 m/s. These facts were not known beforehand and both prove to be valuable information for use in the development of a testing facility and completion of the work at Notre Dame.

#### 4.2 Torsional Hopkinson Bar Testing



c) Split Hopkinson  
Torsional Bar

FIGURE 8

In still more related work, R. Caspar, under the supervision of the author, developed a working clamp design for the torsional Hopkinson bar at AWEF. The torsional Hopkinson bar loads thin walled cylinders dynamically in torsion as shown schematically in Figure 8. This is the same type of testing apparatus used by Duffy<sup>15</sup> and Giovanola<sup>30</sup> in their investigation of shear localization. It was hoped that Caspar could complete the construction of the device by modifying the existing, non-working design of the clamp. After that he was to demonstrate that the apparatus worked by testing a known material. Later he was to test Filler E, an explosive material simulant, in the apparatus. The ultimate goal of Caspar's work is to test explosive materials on a torsional bar of his own design; these short term goals were seen as an excellent beginning toward the ultimate goal.

Caspar was successful in redesigning the clamp and making it work. He enlarged some of the support pins, machined a cylindrical groove in the clamp faces and used sand paper to roughen the surfaces. Furthermore, he redesigned the supports on the clamp so that clamp load was more symmetrically applied. These design modifications led to a working torsional bar. He demonstrated his success by recording the pulse generated by his properly functioning clamp and showing it to be exactly as predicted by simple theory.

Next, Caspar tested steel specimens and got some bewildering results. The stress-strain curves were not as expected from other author's results or from quasistatic tests. Tests were repeated on aluminum instead of steel giving different, but equally unsuccessful, results. After much evaluation of the apparatus, clamp, electronics etc. it was discovered that the specimen was being deformed in the apparatus before the dynamic test. The apparatus was changed to prevent this undesirable deformation and the results were much improved. Copper and aluminum were tested giving good agreement with references. Finally, Filler E was tested resulting in a very low measured plastic strength and fracture strain in the specimen. Further tests will be performed at Notre Dame after a bar is fabricated. More details of this work are reported in Caspar's Graduate Student Summer Research Report, base number 8.

The effort to test explosive materials at Notre Dame will be greatly enhanced by the experience gained at AWEF. Our success in testing Filer E in the torsional Hopkinson bar makes pursuit of torsional testing of explosive materials justified with a high likelihood for success.

## 5. Summary

In the introduction it was mentioned that, by the estimation of the author, the purpose of the Faculty Summer Research Program, in terms of benefiting the faculty participants, is three-fold: to introduce a faculty member to the research interests of the Air Force, to make unique experimental capabilities available to the faculty member and to help the faculty member develop graduate student researchers in the same setting through cooperative use of the Graduate Student Summer Research Program. Each of these objectives was successfully achieved in the tour of the author. A new understanding of the research interests of the Air Force, particularly the Munitions Directorate, was attained as outlined in Section 2. Unique experimental capabilities were used by the author both in his own work and through graduate students. The author was able to use the Cordin 330 camera in conjunction with Fresnel lenses and a macro lens from Pentax to prove the feasibility of recording fracture as it follows shear localization at a notch tip. One graduate student was able to use a 30 mm powder gun, and all the safety equipment associated with such a gun, to launch projectiles up to 500 m/s in punch tests. And, another graduate student was able to use an existing torsional Hopkinson bar to test an explosive simulant and gain valuable experience to be used in building such a bar at Notre Dame. Obviously, from the previous two statements, the author was able to develop graduate student researchers through cooperative use of the Graduate Student Summer Research Program.

## References

1. F.A. Williams, M. Barrere and N.C. Huang (1969), *Fundamental Aspects of Solid Propellant Rockets*, Slough; (Published for) the Advisory Group for Aerospace Research and Development of N.A.T.O. (by) Technivision Services; London; Distributed by Technical Pub.
2. A.W. Campbell, W.C. Davis, J.B. Ramsey and J.R. Travis (1960), "Shock initiation of solid explosives," *J. Phys. Fluids*, **4**, 4, p. 511
3. C.L. Mader (1978), *Numerical Modeling of Detonation*, Univ. California Press, Berkeley, CA
4. E.L. Lee and C.M. Tarver, "Phenomenological model for the shock initiation of heterogeneous explosives," *Phys. Fluids*, **23**, p. 2362
5. W.C. Davis (1981), "High explosives: the interaction of chemistry and mechanics," *Los Alamos Science*, Los Alamos National Lab., Los Alamos, NM, **2**, p 48
6. F.P. Bowden and Y.D. Yoffe (1952), *Initiation and growth of explosions in liquids and solids*, Cambridge University Press, New York
7. P.-A. Persson (1994), *Rock Blasting and Explosives Engineering*, CRC Press, Boca Raton, FL
8. H.W. Leipman and A. Roshko (1957), *Elements of Gas Dynamics*, Wiley and Sons, New York
9. W. Fickett and W.C. Davis (1979), *Detonation*, Univ. California Press, Berkeley, CA
10. Y.C. Fung (1994), *A First Course in Continuum Mechanics*, Third Edition, Prentice Hall, Englewood Cliffs, New Jersey

11. Dyna East Corporation Technical Report DE-TR-91-15 (1991), "Explosive response to unplanned stimuli," Dyna East Corporation, 3201 Arch Street, Philadelphia, PA 19104
12. ASM Handbook (1985), "High strain rate testing," M.R. Staker, Chairman, ASM International, Metals Park, OH, p. 185
13. P.C. Chou and W. Clark (1995), "Blunt cylinder impact for determination of constitutive equations of explosives," *Workshop on Energetic Material Ignition Micromechanics*, July–August 1995, Army Research Lab., Weapons Technology Directorate, Aberdeen Proving Ground, MD
14. R.J. Clifton, J. Duffy, K.A. Hartley and T.G. Shawki (1984), "On critical conditions for shear band formation at high strain rates," *Scripta Metall.* **18**, p. 443
15. A. Marchand and J. Duffy (1988), "An experimental study of the formation of shear bands in a structural steel," *J. Mech. Phys. Sol.*, **36**, 3, p. 251
16. R.C. Glenn and W.C. Leslie (1971), *Metall. Trans.*, **2**, p. 2945
17. A.T. Zehnder and A.J. Rosakis (1991), "On the temperature distribution at the vicinity of dynamically propagating cracks in 4340 steel," *J. Mech. Phys. Sol.*, **26**, p. 151
18. J.J. Mason and A.J. Rosakis, "The Dependence of Dynamic Crack Tip Temperature Fields upon Velocity and Material Parameters," *Mechanics of Materials*, Vol. 16, p 337, 1993
19. M. Zhou, A. Needleman and R.J. Clifton (1994), "Finite element simulations of shear localization in plate impact," *J. Mech. Phys. Solids*, **42**, 3, p. 423
20. M. Zhou, R.J. Clifton and A. Needleman (1992), "Shear band formation in a W–Ni–Fe alloy under plate impact," in *Tungsten and Tungsten Alloys—1992*, Metal Powder Industries Federation, Princeton, NJ, p. 343
21. J.N. Johnson, P.K. Tang and C.A. Forest (1985), "Shock wave initiation of heterogeneous reactive solids," *J. Appl. Physics*, **57**, (9), p 4323
22. P.A. Taylor (1985), "The effects of material microstructure on the shock sensitivity of porous granular explosives," *Proceedings-Eighth Symposium (International) on Detonation*, NSWC MP 86-194, Naval Surface Weapons Center, White Oak, MD, p 358
23. R.E. Setchell, and P.A. Taylor (1984), in *Shock waves, Explosions and Detonation*, AIAA Progress in Astronautics and Aeronautics, J.R. Bowen, N. Manson, A.K. Oppenheim and R.I. Soloukhin, Eds., **95**, AIAA, New York, NY
24. P. Howe, R. Frey, B. Taylor and V. Boyle (1976), *Proceedings-Sixth Symposium (International) on Detonation*, Office of Naval Research-Department of the Navy, Arlington, VA, ACR-221
25. M.M. Carroll and A.C. Holt (1973), "Static and dynamic pore-collapse relations for ductile porous materials," *J. Appl. Physics*, **43**,
26. M.R. Baer and J.W. Nunziato (1986), *Int. J. Multiphase Flow*, **12**, p 861
27. P.B. Butler, M.F. Lembeck and H. Krier (1982), *Comb. Flame*, **46**, p 75
28. J.M. Powers, D.S. Stewart and H. Krier (1990), "Theory of two-phase detonation—Part I: modeling," *Comb. Flame*, **80**, p 264
29. W. Cook (1995), *Private Communication*, Armament Directorate, Eglin AFB, FL
30. J.H. Giovanola (1988), "Adiabatic shear banding under pure shear loading part I: direct observation of strain localization and energy dissipation measurements," *Mech. Mat.*, **7**, p. 59
31. J.F. Kalthoff (1987), Shadow optical analysis of dynamic shear fracture, in: *SPIE Vol. 814 Photomechanics and Speckle Metrology*, 531

32. J.F. Kalthoff and S. Winkler (1987), Failure mode transition at high rates of shear loading, in: C.Y. Chiem, H.-D. Kunze and L.W. Meyer, eds., *Impact Loading and Dynamic Behavior of Materials*, Verlag, Vol. 1, 185
33. J.J. Mason, A.J. Rosakis and G. Ravichandran, "Full Field Measurements of the Dynamic Deformation Field Around a Growing Adiabatic Shear Band at the Tip of a Dynamically Loaded Notch," *J. Mech. Phys. Sol.*, **42**, No. 11, p 1679
34. J. Zimmerman (1995), "Effects of aging treatment and loading geometry upon shear localization in Kalthoff impact tests of C-300 steel," Master's Thesis, University of Notre Dame, Notre Dame, IN
35. J.J. Mason, J. Lambros and A.J. Rosakis (1992), "On the Use of a Coherent Gradient Sensor in Dynamic Mixed-Mode Fracture Mechanics Experiments," *J. Mech. Phys. Sol.*, **40**, No. 3, pp. 641-661
36. R.B. Frey (1981), "The initiation of explosive charges by rapid shear," *Proc. Seventh Int. Symp. on Detonation*, p. 36

# **MOTION-BASED AUTOMATIC TARGET DETECTION, TRACKING, AND RECOGNITION**

**Rajiv Mehrotra**

**Associate Professor**

**Department of Mathematics and Computer Science**

**University of Missouri-St. Louis**

**8001 Natural Bridge Road**

**St. Louis, MO 63121**

**Final Report for**

**Summer Faculty Research Program**

**Wright Laboratory**

**Sponsored by:**

**Air Force Office of Scientific Research**

**Bolling Air Force Base, DC**

**and**

**Wright Laboratory**

**August 1995**

# **MOTION-BASED AUTOMATIC TARGET DETECTION, TRACKING, AND RECOGNITION**

**Rajiv Mehrotra**

**Associate Professor**

**Department of Mathematics and Computer Science**

**University of Missouri-St. Louis**

## **Abstract**

Motion is an important cue used by several biological vision systems for object recognition and scene analysis. Static image feature-based machine vision systems for target detection and recognition usually fail in case of camouflaging, certain climatic and scene illumination that yield poor contrast images, and for certain types of images such as IR images. In such cases, techniques that utilize motion-based techniques are expected to yield better performance. Several experimental studies in vision psychology have demonstrated human visual system's capability to recognition/classification object based on the motion trajectory of a selected set of feature points. It is evident from these studies that moving point trajectories carry object shape related information. This project is concerned with the study of issues pertinent to motion trajectory-based automatic target detection, tracking, and recognition with an objective of developing founding technologies. This research was initiated as part of my Summer AFOSR Faculty Research Associateship at Wright Laboratory. The findings of my initial study of important technical issues, current accomplishments, and proposed future research are described in this report.

## **1. Introduction**

Motion plays a crucial role in biological vision systems. Humans have ability to recognize a distant person, whose face is not clearly visible, by his/her motion characteristics, such as walk, gait, hand gestures, etc. In animal kingdom, the camouflage strategies used by preys and predators utilize the fact that motionless objects in a natural scenes are often difficult to detect and recognize. Dynamic changes in a scene cause the focus of the attention of the observer(s) on moving objects. Several clinical studies of biological visual systems and research in vision and cognitive psychology have discovered remarkable motion perception and interpretation capabilities of biological visual systems. As a result, over the last decade, there has been a great deal of interest in developing computer vision systems capable of detecting motion characteristics from a sequence of images of a dynamic scene and utilizing detected motion characteristics to analyze and interpret scenes.

Model-based techniques are commonly used for automatic target (or object) recognition. Model-based ATR involves extraction of low-level features from the input image(s), building scene description, and matching the scene description against the precompiled models of the targets to find the matching target. Existing computer vision techniques for target (or object) recognition techniques utilize intensity boundary or surface-based features for building target models and unknown target description. There are several cases where the scene or image characteristics are such that intensity-based features cannot be reliably detected and hence target recognition solely based on intensity features is impossible. Such cases include images of camouflaged targets. In such cases, motion-based (or temporal) features can be utilized for automatic target detection, tracking, and recognition. Moving or temporal edges, temporal corners, discontinuities in optical flow, optical flow (or temporal) fields and patterns, moving point trajectories are examples of temporal features. The goal of this project is to investigate the role of motion trajectories in automatic target detection, tracking, and recognition.

## **2. Motivations for Motion-Based ATR**

In psychology, Johansson's moving light display (MLD) has been extensively used to study motion perception capabilities of human visual system [1, 2, 3]. MLDs consists of light bulbs or bright spots attached to a few selected locations of one or more completely dark objects (e.g., joints of humans dresses in black or corners of a black colored cube) that move in front of a dark background. The set of static bright spots belonging to an object do not carry any structural information and therefore meaningless to observers. On the other hand, a light display of moving objects (e.g., rotating cube; walking, running or dancing humans) yields

meaningful interpretations such as the identity of the shape, gender of a person or even the gait of a friend, etc. These studies indicate that we use motion information for object recognition.

There are two theories have been proposed to explain these motion-based recognition capabilities of the human visual system. The first theory advocates that the motion information is used to recover the 3-dimensional structure, and then use the structure for recognition. According to the second theory, motion information is directly used, without the recovery of structure, for object recognition. In recent years there has been a growing interest in the computer vision community, in the development of computational models based on both of these theories. The investigation of motion-based ATR is motivated by desire to develop ATR techniques that can duplicate the motion-based recognition capability of the human visual.

Computer vision approaches for structure from motion, that recover 3-dimensional coordinates of moving object points and their 3-dimensional motion from a sequence of image frames [4, 5], are based on the first theory. These approaches assume that recovered structural information will be used for recognition. However, surface recovery is very sensitive to noise and it has been found that a reliable recognition is usually not possible solely based on the the recovered structural information.

The computational approaches based on the second theory of motion-based recognition involve the direct use of motion information extracted from a sequence of images for object and/or their motion recognition [6, 7, 8]. It was found in a study that inverted (upside down) MLDs are not usually recognized [9]. This suggests that the familiarity with a particular motion plays a key role in motion-based recognition. Since an inverted motion is not a familiar or natural motion, it is difficult to recognize. This implies that temporal feature-based object (target) models need to be employed for motion-based object (target) recognition. It is also clear from the MLD studies that in general, motion trajectories of a few selected points (e.g., joints, endpoints of selected limbs, etc.) over a large sequence of images are usually sufficient for object recognition. The optimal set of object points that need to be tracked over time for a successful recognition is object shape dependent.

Like the traditional model-based recognition, the motion-based object (or motion) recognition/classification involves two phases: the *model-building* phase and the *recognition* phase. In the model-building phase, temporal feature-based object models of the known objects (or motion) are built and in the recognition phase the description of unknown object (or



motion) extracted from an input image sequence is matched with a model. One can also employ the traditional static feature-based models, but utilize motion information to obtain a robust and reliable segmentation of a frame of the input sequence. The type of features employed in models, reliability of their extraction, and their noise robustness have a direct influence on the overall accuracy of any motion-based recognition approach.

### 3. Motion-Based Object or Scene Modeling

For motion-based object recognition, objects and their motions need to be modeled in terms of their temporal or motion-based features. Object features need to be extracted from input image sequences. For example, the motion trajectories of a set of points need to be detected from an input image sequence for an MLD type object recognition system. Moving object boundary or region-based features can be used for object modeling. In other words, input image sequence needs to be processed (or segmented) to detect the desired set of features. There are two major classes of temporal features:

(i) low-level features that can be detected directly from an image sequence – examples include moving edges, moving corners, and moving textures.

(ii) features based on the motion (i.e., optical flow or point motion trajectories) extracted from the input image sequence – examples include optical flow-based features and motion trajectory-based features.

Moving edges can be detected by combining temporal and spatial gradient [10]. In other words, the local maxima of  $|E_k(i, j)| \cdot D_k(i, j)$  are moving edge points. Here,  $E_k(i, j)$  is the spatial gradient at point  $(i, j)$  of frame  $K$ , and  $D_k(i, j) = |F_k(i, j) - F_{k-1}(i, j)|$  is the magnitude of the intensity difference between frame  $k$  and  $k-1$  at point  $(i, j)$ . Similarly, the moving corners can be detected by the product of cornerness  $C_k(i, j)$  and  $D_k(i, j)$ . Any static corner detector can be used to compute the cornerness.

Alternatively, moving edges can be found by locating the zero-crossings of the convolution of the intensity history at each point with the second derivative in time of the temporal Gaussian smoothing function [11], i.e., the spatial zero-crossings of  $(\partial^2 G(t)/\partial t^2) * V$ , where  $G(t) = (s/\sqrt{\pi}) \exp(-s^2 t^2)$ ,  $V$  is the vectors of intensity history at a point,  $*$  denotes the convolution in time, and  $s$  is a constant.

The discontinuities in the optical flow field correspond to the boundary points of moving rigid objects [12–14]. Optical flow discontinuity can be detected either by integrating discontinuity detection into the optical flow computation model [12, 13] or by applying discontinuity any detection approach (e.g., LoG operator) to the optical flow data estimated from a sequence of image [15]. Temporal textural can be represented by statistical measure such as average and variance of the optical flow over a region or by co-occurrence matrix-based measures computed from the optical flow field [16].

Features based on the trajectories of moving points, such as trajectory segments, discontinuities, curvature as a function of time, relative motion of two or more points, etc. can be used for motion-based object modeling. It is evident that in the case of MLDs, information present in moving point trajectories is utilized for object recognition. Given a large sequence of images of a dynamic scene, a critical task in any trajectory-feature-based computational approach to scene analysis is to obtain, from the input sequence, the trajectory set associated with a set of points belonging to the moving objects. Optical flow methods yield dense velocity field (i.e., velocity vector at each image point) and can be used to obtain the motion trajectories of a selected set of points. Unfortunately, optical flow method provide far more than needed information and are computationally intensive, specially for a large image sequence. Therefore, optical flow method are not well suited for moving point trajectory detection. Alternatively, a set of temporal feature points (e.g., temporal edge, corner points, or object points) can be detected in each of the images of a sequence and correspondence of the selected points over the given sequence can be established to obtain their motion trajectories. Constraints such as rigidity or trajectory smoothness can be utilized to establish point correspondences.

The primary objective of this research to develop founding technologies for motion trajectory-based automatic target recognition. This includes investigation of issues pertinent to motion-based target recognition and development of reliable and efficient techniques for motion trajectory-based target detection, tracking, and recognition. Specifically, the major research issues include:

1 Detection of point motion trajectories over a large monocular image sequence – Given a set of  $m$  points over  $n$  frames, a total of  $m^{(n-1)}$  trajectory sets are possible. Of these, only  $m$  trajectories are the correct ones. For  $m = 5$  and  $n = 4$ , the total number trajectory sets is  $(5!)^3 = 1,728,000$ . It is obvious (i) that even for a moderate size point set and image sequence,

it is impossible to try all possible trajectory sets and (ii) that a method is needed to identify the correct trajectory set. The problem becomes more complex when trajectories can terminate or start at intermediate frames of a sequence. In this case, the point set size is not equal for all the frame. I have developed and implemented a heuristic for the detection of motion trajectories of points belonging objects in the scene. The performance of this algorithms is currently being empirically evaluated. Initial experimental results are quite encouraging.

2 Motion trajectory-based automatic target tracking – One approach to target tracking is to identify all moving regions in each of the frames of a sequence and select feature points for each region. The trajectory of points belonging to each regions track the corresponding object. In some cases, a meaningful and reliable image segmentation may not be possible. In such cases, to track moving objects in a scene, the detected trajectory set need to be partitioned into clusters belonging to different objects. The point or trajectory clustering can be done for each pair of adjacent frames after the point correspondence is established. We are currently investigating such issues and developing trajectory-based automatic target tracking techniques. The principles underlying these approaches are briefly discussed in section 5.

3. Motion trajectory-based target recognition: MLD studies have established that point trajectories-based object recognition is possible for a large class of object. Can a computer vision technique be developed for trajectory-based automatic target recognition? We propose to investigate this issue in the next phase of this research. It is our belief that point trajectory-based model can be developed to represent targets and for trajectory-model-based object recognition. Our research plans include study two types of trajectory-based models:

- Individual point trajectory feature-based models – This approach model an object as the set of trajectories of its selected feature points and each trajectory is represented by its features or parameters. For example, a simple model for the side-view of a moving car can be the set of trajectories of 6 points (1 point on each of the two visible wheels, one point on the rear of the body, one on the front of the body, and the front and the rear points of the top of the body). Assuming that the car moves on a planar surface, this simple model consists of two circular trajectories corresponding to the motion of points on each of the two visible wheels, and four linear trajectories corresponding to the points on the body of the car. This modeling approach can be used for trajectory-based identification (or classification) of a large class of objects and motion type. We propose to study such modeling techniques and develop corresponding target recognition methods.

- Models based on relative motion of features points – For more general and robust target recognition, modeling techniques that represent an object in terms of relationships among point trajectories appear to be more useful. For example, the simple car model can be generalized by substituting the 4 body point linear trajectories by the facts that the front and rear body point trajectories always are nearly the same (i.e., have same parameters) and the trajectories of the front and rear end body-top points are nearly the same. The assumption of motion on a planar surface can be eliminated in case of this extended car model. The relationships among trajectories of moving points carries important structural information. We propose to study the effective of this modeling approach in automatic target recognition.

The ultimate and long term objective of this research is to build founding technologies and techniques for integrated, generalized, efficient, and reliable automatic trajectory-based target detection, tracking, and recognition. The trajectory detection and target tracking are the topics of immediate focus in this project. Some important results of our research on these topics and proposed future activities are discussed in the following sections.

#### 4. Trajectory Detection Technique

Consider a sequence of  $n$  images (frames),  $S = \{f_1, f_2, \dots, f_n\}$ . Assuming that a set of feature points have already been detected in each of the frames, a frame is a set of points, i.e.,  $f_i = \{p_i^1, p_i^2, \dots, p_i^m\}$ , where  $p_i^j$  is the 2-dimensional coordinates of the  $j$ -th point of the  $i$ -th frame and the value of  $m$  can be different for different frames. The objective is to find the establish motion correspondence of frame points to find their trajectories.

Our approach to trajectory detection is based on the assumptions that the trajectory of each moving point is as "smooth" as possible and the length of the trajectory of a point should be as "small" as possible. The assumption of smoothest and shortest length trajectory is motivated by the principle of inertia of motion. The motion characteristics of a physical entity doesn't change instantaneously, due to inertia. Consequently, the trajectory of every moving point is smooth. The trajectory smoothness concept is also supported by the principle of "visual inertia" in vision psychology [17] which suggests that when any object moves in one direction at uniform velocity we tend to perceive it as continuing its motion in that direction. The assumption of short length trajectory is strongly supported by the vision psychology principle of least action [18], which suggests that when we perceive a moving object, we tend to perceive it as moving

along a path that in some sense is the shortest, simplest, and most direct.

Since in real scenes, occlusions/disocclusions of moving objects are quite common, we cannot assume that the same number of points appear in all the frames of a sequence and that for every point there is a corresponding point in all the frames. In other words, an ideal approach must permit termination of exiting point trajectories and beginning of new trajectories at intermediate frames within the given frame sequence. We formulate the trajectory detection problem as an optimization problem. The cost function to be optimized comprises of the following terms:

- **Trajectory Smoothness** – For a point motion trajectory to be smooth, its velocity (and hence the displacement) should be small (or relatively unchanged) over a small time interval (i.e., between adjacent pairs of frames). The velocity (displacement) is a vector and therefore both the velocity (displacement) direction and the velocity (displacement) magnitude should be relatively unchanged.

We are experimenting with the following measures for the trajectory smoothness (defined over three adjacent frames  $t-1$ ,  $t$ , and  $t+1$ ) to evaluate their relative effectiveness for different types of motion.

**Measure 1:** This measure is the sum of the displacement direction and the displacement magnitude smoothness measures:

$$0.5 \left[ 0.5 \left[ 1 + \frac{\overrightarrow{p_{t-1}^i p_t^j} \cdot \overrightarrow{p_t^j p_{t+1}^k}}{|\overrightarrow{p_{t-1}^i p_t^j}| \cdot |\overrightarrow{p_t^j p_{t+1}^k}|} \right] + \left[ 1 + \frac{|\overrightarrow{p_{t-1}^i p_t^j}| - |\overrightarrow{p_t^j p_{t+1}^k}|}{\max_{m,n,r} (|\overrightarrow{p_{t-1}^m p_t^n}| - |\overrightarrow{p_t^n p_{t+1}^r}|)} \right] \right]$$

where  $m$  belongs to  $t_{t-1}$ ,  $n$  belongs to  $t_t$ ,  $r$  belongs to  $t_{t+1}$ . The first term is the direction smoothness term. The value of the direction smoothness measure ranges between 0 and 1. The minimum value of 0 is attained when the angle between the vectors is  $180^\circ$  and the maximum value of 1 is attained when the angle between the vector is  $0^\circ$ . The second term is the displacement magnitude smoothness. Its value ranges between 0 and 1. The value of 0 is for the worst trajectory and it is 1 for the best trajectory. These two terms can be weighed differently to emphasize the relative importance of the two components (i.e., direction and magnitude

smoothness).

**Measure 2:** This measure utilizes the magnitude of the difference between the vectors  $\overrightarrow{p_{t-1}^i p_t^j}$  and  $\overrightarrow{p_t^j p_{t+1}^k}$  to define the trajectory smoothness as

$$\left[ 1 - \frac{|\overrightarrow{p_{t-1}^i p_t^j} - \overrightarrow{p_t^j p_{t+1}^k}|}{\max_{m,n,r} |\overrightarrow{p_{t-1}^m p_t^n} - \overrightarrow{p_t^n p_{t+1}^r}|} \right]$$

where  $m$  belongs to  $f_{t-1}$ ,  $n$  belongs to  $f_t$ ,  $r$  belongs to  $f_{t+1}$ . The maximum value is 1 for the best trajectory and the minimum value is 0 for the worst trajectory.

**Measure 3:** Transform the 2D vector  $\overrightarrow{p_{t-1}^i p_t^j} = \begin{bmatrix} dx \\ dy \end{bmatrix}$  into a 3D vector as  $\overrightarrow{v_{t-1}^i v_t^j} = \begin{bmatrix} dx \\ dy \\ 1 \end{bmatrix}$ .

For two adjacent 3D displacement vectors  $\overrightarrow{v_{t-1}^i v_t^j}$  and  $\overrightarrow{v_{t-1}^j v_{t+1}^k}$ , the trajectory smoothness =

$$0.5 \left[ 1 + \frac{\overrightarrow{v_{t-1}^i v_t^j} \cdot \overrightarrow{v_{t-1}^j v_{t+1}^k}}{|\overrightarrow{v_{t-1}^i v_t^j}| \cdot |\overrightarrow{v_{t-1}^j v_{t+1}^k}|} \right]$$

The value of this measure ranges between 0 (for worst trajectory) and 1 (for best trajectory).

• **Trajectory Length** – For a point trajectory to be of the minimum possible length, the displacement of the point between the adjacent frames should be as small as possible. The two different measures that we are employing in our experiments are:

**Measure 1:**  $\left[ 1 - \frac{|\overrightarrow{p_{t-1}^i p_t^j}| + |\overrightarrow{p_t^j p_{t+1}^k}|}{\max_{m,n,r} (|\overrightarrow{p_{t-1}^m p_t^n}| + |\overrightarrow{p_t^n p_{t+1}^r}|)} \right]$  and

**Measure 2:**  $0.5 \left[ \left[ 1 - \frac{|\overrightarrow{p_{t-1}^i p_t^j}|}{\max_{m,n} (|\overrightarrow{p_{t-1}^m p_t^n}|)} \right] + \left[ 1 - \frac{|\overrightarrow{p_t^j p_{t+1}^k}|}{\max_{r,s} (|\overrightarrow{p_{t-1}^r p_t^s}|)} \right] \right]$

In measure 1 above,  $m$  belongs to  $f_{t-1}$ ,  $n$  belongs to  $f_t$ ,  $r$  belongs to  $f_{t+1}$ , whereas in measure 2, where  $m$  belongs to  $f_{t-1}$ ,  $n$  and  $r$  belong to  $f_t$ , and  $s$  belongs to  $f_{t+1}$ . Both these measures have the minimum value of 0 for the worst trajectory and the maximum value of 1 for the best trajectory.

#### 4.1. Optimization Problem

The quality a trajectory  $p_{t-1}^i p_t^j p_{t+1}^k$  over the three adjacent frames  $f_{t-1}$ ,  $f_t$ ,  $f_{t+1}$  is the sum of the corresponding trajectory smoothness and length measures:

$$T(p_{t-1}^i p_t^j p_{t+1}^k) = w_1 S(p_{t-1}^i p_t^j p_{t+1}^k) + w_2 L(p_{t-1}^i p_t^j p_{t+1}^k)$$

where  $T(a)$ ,  $S(a)$ , and  $L(a)$  denote the trajectory quality, the trajectory smoothness measure, and the trajectory length measure associated with the 3-point trajectory  $a$ , respectively.  $w_1$  and  $w_2$  are the relative weight constants.

The objective is to find the trajectory set that optimizes the sum of the trajectory quality over all the frames in the sequence. Thus, the optimization problem is to find trajectory set that maximizes

$$\sum_{t=2}^{n-1} \sum_{i=1}^{|f_{t-1}|} \sum_{j=1}^{|f_t|} \sum_{k=1}^{|f_{t+1}|} T(p_{t-1}^i p_t^j p_{t+1}^k) \quad (1)$$

under the constraint that every 3-point segment of each trajectory is a valid segment with respect to the motion uniformity constraint, i.e.,

$$T(p_{t-1}^i p_t^j p_{t+1}^k) > Q$$

where  $Q$  is a predefined threshold,  $t = 2, \dots, n-1$ ,  $i = 1, \dots, |f_{t-1}|$ ,  $j = 1, \dots, |f_t|$ , and  $k = 1, \dots, |f_{t+1}|$ . Note that a trajectory segment's validity can be also determined by using separate thresholds for the smoothness and length measures.

#### 4.2. Algorithm

In this section, we present a polynomial time heuristic to obtain a near optimal solution to the above optimization problem. For each triplet of adjacent frames, the trajectory set is updated by finding the 3-point trajectory segments that are valid and that either define the best extensions of existing trajectories or beginning of new trajectories. The process of finding

locally best trajectories is repeated for every set of three adjacent frames of the given sequence. Earlier points trajectory detection approaches [19, 20] either assume that point trajectory over the first two frames are given and extend those initial trajectories over the following frames of the sequence or use time consuming multiple forward and backward iterations over the entire frame sequence to find the trajectory set. The technique presented below doesn't have any of these limitations. It assumes that a point can belong to at most one trajectory. The steps of the algorithm are:

Step 1: The current trajectory set  $\Psi = \phi$ . For  $t = 2$  to  $n-2$  repeat Steps 2 and 3.

Step 2: Let  $\Gamma$  be the set of all the 3-point trajectories in frame triplets  $\{f_{t-1}, f_t, f_{t+1}\}$ . Compute the motion uniformity  $T$  for all 3-point trajectories in  $\Gamma$  that either consists of points that are not assigned to any trajectory in  $\Psi$  or contains a point of frame  $f_{t+1}$  which extends an existing trajectory. Discard all trajectories  $a$  with  $T(a) > Q$  ( $Q$  a threshold) from further consideration. Let  $\Theta$  be the set of trajectories found in this step.

Step 3: Repeat this step until  $\Theta$  is empty. Find the best trajectory(ies) in  $\Theta$  (i.e., trajectory(ies) with maximum value of  $T$ ). Let  $\zeta$  be the set of best trajectories. If two or more conflicting best trajectories (i.e., they share one or more points) are found and  $t < n-2$ , go to Step 3A. Otherwise, go to Step 3B.

Step 3A: For each 3-point trajectory  $\tau = (p_{t-1}^i p_t^j p_{t+1}^k)$  in  $\zeta$ , compute its extendibility  $E(\tau)$  using the following definition of trajectory extendibility:

A trajectory  $\tau = (p_{t-1}^i p_t^j p_{t+1}^k)$  is considered extendible to a point  $x$  of frame  $f_{t+2}$ , if the 3-point trajectory segment  $(p_t^j p_{t+1}^k x)$  is

- (i) valid,
- (ii) the best of all 3-point trajectories that contain  $p_t^j$  and  $p_{t+1}^k$  as the first two points and any point of the frame  $f_{t+2}$  as the third point, and
- (iii) the best of all 3-point trajectories constaining  $x$  as the last point and any points of the previous two frames  $f_t$  and  $f_{t+1}$ .

The extendibility of  $\tau$  is measure by

$$E(\tau) = T(\tau) + T(p_t^j p_{t+1}^k p_{t+2}^z)$$



where  $T(p_t^j p_{t+1}^k p_{t+2}^z) > Q$ ,

$$T(p_t^j p_{t+1}^k p_{t+2}^z) = \max_{p_{t+2}^u \text{ in } f_{t+2}} T(p_t^j p_{t+1}^k p_{t+2}^u) \text{ and}$$

$$T(p_t^j p_{t+1}^k p_{t+2}^z) \geq \max_{p_t^j \text{ in } f_t \text{ and } p_{t+1}^k \text{ in } f_{t+1}} T(p_t^j p_{t+1}^k p_{t+2}^z).$$

Find the best extendible trajectory(ies). The set  $\varsigma$  = the set of best extendible trajectory(ies).

Step 3B: Apply the appropriate case:

Case 1:  $|\varsigma| = 1$ : Let  $\tau_b$  be the trajectory in  $\varsigma$ . If the first two points of  $\tau_b$  don't belong to any trajectory in  $\Psi$ , include  $\tau_b$  as a new trajectory in  $\Psi$ ; Otherwise, extend an existing trajectory in  $\Psi$  that has the same last two points as the first two points of  $\tau_b$  to the last point of  $\tau_b$ . Remove  $\tau_b$  and all trajectories containing one or more points of  $\tau_b$  from  $\Theta$ .

Case 2:  $|\varsigma| > 1$ : Randomly select any trajectory  $\tau_b$  from  $\varsigma$ . If the first two points of  $\tau_b$  don't belong to any trajectory in  $\Psi$ , include  $\tau_b$  as a new trajectory in  $\Psi$ ; Otherwise, extend an existing trajectory in  $\Psi$  that has the same last two points as the first two points of  $\tau_b$  to the last point of  $\tau_b$ . Remove  $\tau_b$  and all trajectories containing one or more points of  $\tau_b$  from  $\Theta$ .

The above algorithm ensures that for each triplet of adjacent frames, the set of existing trajectories is updated by including only the locally best trajectories. Conflicts are resolved by using the notion of trajectory extendibility. In other words, a one frame look ahead is employed to determine the best trajectories in case of conflicts. We found that Case 2 of Step 3B is true only when a point is shared by two or more actual trajectories. We plan to extend this algorithm to handle such exceptional cases as well as to handle dynamic scenes where points may disappear and reappear at some later frames. We have conducted some tests with this algorithm using a few small size data sets (a maximum of 8 frame sequence, with at most 10 points per frame). The initial results are very encouraging [26]. The performance of the algorithm varies with different measures for the trajectory smoothness and the trajectory length criteria proposed in section 4. Extensive testing and performance evaluation with large data sets need to be conducted to evaluate the overall accuracy of the algorithm as well as the relative effectiveness of various

trajectory smoothness and length measures for different motion types and frame sequence characteristics.

## **5. Feature-Point-Based Target Tracking**

I am currently developing techniques for trajectory-model-based rigid target tracking. The fundamental ideas behind these techniques are briefly described in this section.

**Approach 1:** The ability to track a moving target is common to many biological vision systems. In biological vision systems, on detecting a moving stimulus, the object of interest is brought near the fovea by a saccadic eye movement. Then the pursuit of that object starts. The moving target is stabilized on the fovea by smooth eye movements. Thus, the target tracking process consists of two distinct steps: (i) detection and discrimination of moving targets and (ii) then pursuit of the object of interest.

Major steps of a automatic target tracking technique (currently under development), which is based on this two step process are outlined below:

**Step 1:** Perform motion-based segmentation of frames of an input image sequence to detect regions corresponding to targets moving relative to the camera. As mentioned in section 3, several temporal region and edge-based segmentation methods already exist that can be employed for this purpose.

**Step 2:** For each detected region of a frame, a set of feature (or interest) points is selected. Examples of such feature points include, boundary points with high corneriness, points with high intensity variance within a region, boundary points with locally maximal curvature, vertices of the polygonal approximation of a region boundary, etc.

**Step 3:** Employ the technique proposed in section 4 to find motion trajectory associated with each of the selected points.

**Step 4:** Trajectories of points belonging to a moving region track the movement of that object and its component. If desired, the representative trajectory associated with the image region of a rigid object can be computed for any number of adjacent frames from the detected trajectories of its points over the selected frame set.

**Approach 2:** The accuracy of the above target tracking method depends on the reliability of the frame segmentation approach. In defense-oriented applications, there are several factors such

as camouflaging, poor scene illumination (due to climatic conditions or night time), or image characteristic (e.g., IR images) that make a reliable or robust segmentation almost impossible. In such cases, a target tracking method that do not heavily rely on the initial segmentation for tracking is desirable.

The major steps of an approach to point-trajectory-based target tracking without a robust initial segmentation are as follows:

Step 1: Utilize a feature point selection approach that doesn't rely on frame segmentation in to moving regions. Examples of such feature points include temporal edge or corner points with some prespecified characteristics, response of Moravec's operator [21], etc..

Step 2: Repeat the following steps for every pair of adjacent frames.

Step 3: Establish point correspondence between the current pair of adjacent frames using the trajectory detection method.

Step 4: Assuming a transformation model (e.g., affine transformation), a group of points that are undergoing similar transformation and are spatially closer to each other can be consider to belong to the same object.

We utilize an affine model. The relationship between the transformation of an object points and their local 2D motion under affine model is derived below:

Assuming that all targets are rigid and that an affine model characterizes the transformation of an object region between a pair of adjacent frames, every feature point  $p(t+1)$  in frame  $f_{t+1}$  is the result of an affine transformation  $[A(t), a(t)]$  of the corresponding point  $p(t)$  in frame  $f_t$ . That is,

$$p(t+1) = A(t)p(t) + a(t) \quad (2)$$

Let the local 2D motion vector associated with all the feature points  $p(t)$  belonging to the image region  $R$  of a moving object in frame  $f_t$  is modeled by an affine model:

$$\forall p(t) \text{ of } R, \dot{p}(t) = B(t)p(t) + b(t) \quad (3)$$

The  $2 \times 2$  matrix  $B(t)$  represents a large class of motion including, rotation, scaling, shear, etc. Affine model of motion field have also been used by other researchers [22-25]. Taylor series expansion of  $p(t+1)$  about  $p(t)$  after dropping the higher order terms yields,

$$p(t+1) = p(t) + \delta t \dot{p}(t) \quad (4)$$

where  $\delta t$  is the time between two successive frames. From (3) and (4), we get

$$p(t+1) = p(t) + \delta t B(t)p(t) + \delta t b(t) \quad (5)$$

Comparing (2) and (5),

$$A(t) = I + \delta t B(t) \quad \text{and} \quad a(t) = \delta t b(t) \quad (6)$$

where  $I$  is a  $2 \times 2$  identity matrix. Thus, the estimation of  $[A(t), a(t)]$  reduces to the estimation of  $[B(t), b(t)]$  characterizing the local 2D motion.

The point correspondences established by the trajectory detection method can be used to estimate the parameters of both affine transformations  $[A(t), a(t)]$  and  $[B(t), b(t)]$ . It is obvious that for a reliable estimation of the transformation parameters associated with the points of an object region, all points belonging to that object need to be identified. This creates a catch 22 situation as to partition the feature points into sets corresponding to different moving objects requires transformation parameters to be known and to estimate the objects' transformation parameters, the point sets corresponding to each of the moving objects need to be known.

One possible approach to get around this problem is to first use the known correspondences of a very few spatially close points to estimate the corresponding transformation parameter. Then cluster those point of a frame together for which the corresponding point in the other frame obtained by the trajectory detection method satisfies the estimated transformation. Transformation parameters estimated by points belonging to different objects, will usually not add any additional points to the clusters. Such transformations should be abandoned. The process of estimating transformation and point clustering can be repeated until all points are grouped into verifiable clusters. Since a transformation estimation step requires solution to the corresponding least square error problem, this process of point clustering is computationally very expensive. An alternate approach to point clustering that doesn't require the estimation of transformation parameters is describe below.

**Point Clustering Without Transformation:** To perform point clustering without knowing the associated transformation parameters, the point correspondence between adjacent frames is used to establish equivalent affine invariant spaces in the two frames. Any two points that undergo the same affine transformation have the same representation in two frames with

respect to the corresponding affine invariant spaces. Corresponding points with invariant representation in the two affine invariant spaces are clustered together to form the point set corresponding to an object.

In a frame, an affine invariant space is defined by a triplet of non-colinear points that are spatially close to each other. Specifically, let  $p_{t0}$ ,  $p_{t1}$ , and  $p_{t2}$  be a triplet of points. The vectors  $p_x = p_{t1} - p_{t0}$  and  $p_y = p_{t2} - p_{t0}$  are linearly independent and hence define a 2D linear basis. As shown in Figure 1, any point  $p$  can be represented in this basis by a pair of scalars  $(\alpha, \beta)$ , such that  $p = \alpha p_x + \beta p_y + p_{t0} = \alpha(p_{t1} - p_{t0}) + \beta(p_{t2} - p_{t0}) + p_{t0}$ .

An application of affine transformation  $T$  transforms the point  $p$  to:

$$Tp = \alpha(Tp_{t1} - Tp_{t0}) + \beta(Tp_{t2} - Tp_{t0}) + Tp_{t0}.$$

Hence, the transformed point  $Tp$  has the same coordinates  $(\alpha, \beta)$  in the transformed basis triplet  $(Tp_{t0}, Tp_{t1}, Tp_{t2})$ . The point corresponding to point  $p$  in the adjacent frame is  $Tp$  and the three points corresponding to the basis triplet points define the transformed basis triplet. The basis triplet points can be normalized to have the coordinates  $(0, 0)$ ,  $(1, 0)$ , and  $(0, 1)$ , respectively.

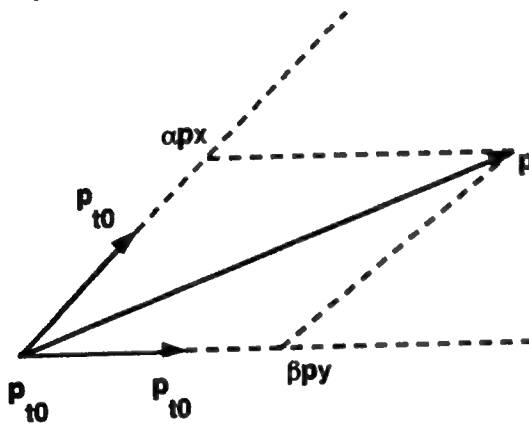


Figure 1: Representation of point  $p$  in the affine basis triplet

## 6. Conclusions

Trajectories of selected points of a moving object carry object shape information. The essential steps involved in any machine vision technique for trajectory-based automatic target detection, tracking, and recognition include (i) feature point selection from each frame of an input image sequence, (ii) detection of motion trajectories of selected feature points, and (iii) trajectory feature-based model-based target recognition. Important issues are being investigated, founding technologies are being developed and tested. Current accomplishments and future research are outlined in this report. Specifically, a new techniques for finding motion trajectories is proposed. Basic principles underlying a couple of techniques for feature point trajectory-based target tracking are described. Trajectory-based target modeling and target recognition concepts are briefly discussed.

## References

- 1 G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception and Psychophysics*, Vol. 14, No. 2, 1973, pp. 210-211.
- 2 G. Johansson, "Visual motion perception," *Scientific American*, June 1975, pp. 76-88.
- 3 C. D. Barlay, J. E. Cutting, and L. T. Kozlowski, "Temporal and spatial factor in gait perception that influence gender recognition," *Perception and Psychophysics*, Vol. 23 No. 2, 1978, pp. 145-152.
- 4 M. Subbarao, "Interpretation of image motion fields: a spatio-temporal approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 11, No. 3, 1989, pp. 266-278.
- 5 T. J. Broida and R. Chellapa, "Estimating the kinematics and structure of rigid objects from a sequence of monocular images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 13, No. 6, 1991, pp. 597-613.
- 6 N. H. Goddard, *The perception of articulated motion: recognizing moving light displays*, PhD dissertation, University of Rochester, 1992.
- 7 J. K. Tsotsos, J. Mylopoulos, H. D. Covey, and S. W. Zuker, "A framework for visual motion understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 2, No. 6, 1980, pp. 563-573.
- 8 D. Koller, N. Heinze, and H-H Nagel, "Algorithmic characterization of vehicle trajectories from image sequences by motion verbs," *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, Maui, HI, 1991, pp. 90-95.
- 9 S. Sumi, "Upside-down presentation of the Johansson's moving light-spot pattern," *Perception*, Vol. 13, 1984, pp. 283-286.

- 10 S. Haynes and R. C. Jain, "Time varying edge detection," *CVGIP*, Vol. 21, 1983, pp. 345-367.
- 11 J. H. Duncan and T-C Chou, "On the detection of motion and the computation of optical flow," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 14, No. 3, 1992, pp. 346-352.
- 12 K. Chaudhury and R. Mehrotra, "Optical flow estimation using smoothness of brightness trajectories," *CVGIP: Image Understanding*, Vol. 60, No. 2, 1994, pp. 230-244.
- 13 K. Chaudhury and R. Mehrotra, "Trajectory-based computational model for optical flow estimation," *IEEE Transactions on Robotics and Automation*, October 1995, To Appear.
- 14 R. Srinivasan, D. Lavine, and L. Kanal, *Model-based ATR using image motion cues from FLIR image sequences*, Wright Laboratory Report WL-TR-93-1110, September 1993.
- 15 W. B. Thompson and T. C. Pong, "Detecting moving objects," *International Journal of Computer Vision*, Vol. 4, 1990, pp. 39-57.
- 16 R. C. Nelson and R. Polana, "Qualitative recognition of motion using temporal texture," *CVGIP: Image Understanding*, Vol. 56, No. 1, 1992, pp. 78-89.
- 17 V. S. Ramachandran and S. M. Antis, "Exploration of motion path in human visual perception," *Vision Research*, Vol. 23, 1983, pp. 83-85.
- 18 R. Shepard and L. Cooper, *Mental images and their transformations*, Cambridge, MA, MIT Press, 1982.
- 19 I. K. Sethi and R. C. Jain, "Finding trajectories of feature points in a monocular image sequence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 9, No. 1, 1987, pp. 56-73.
- 20 K. Rangarajan and M. Shah, "Establishing motion correspondence," *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, Lahania, Hawaii, 1991, pp. 103-108.
- 21 H. P. Moravec, *Robot Rover Visual Navigation*, UMI Research, 1981.
- 22 P. J. Burt, J. R. Bergen, et. al., "Object tracking with a moving camera," *Proceedings IEEE Workshop on Visual Motion*, Irvine, CA, March 1989, pp. 2-12.
- 23 S. Nagahdaripour and S. Lee, "Motion recovery from image sequence using only first-order optical flow information," *International Journal of Computer Vision*, Vol. 9, No. 3, 1992, pp. 163-184.
- 24 F. G. Meyer and P. Bouthemy, "Estimation of time-to-collision maps from first-order motion models and normal flows," *Proceedings 11th International Conference on Pattern Recognition*, Hague, 1992, pp. 78-82.

25 F. G. Meyer and P. Bouthemy, "Region-based tracking using affine motion models in long image sequences," *CVGIP: Image Understanding*, Vol. 60, No. 2, 1994, pp. 119-140.

26 R. Mehrotra, "Finding motion trajectories of feature point," *1996 IEEE International Conference on Robotics and Automation*, Submitted.



**A REVIEW OF ELASTOMERIC SYSTEMS MADE CONDUCTIVE  
THROUGH THE USE OF CONDUCTIVE PARTICLES**

Douglas J. Miller, PhD  
Associate Professor of Chemistry  
Science & Mathematics Department

Cedarville College  
PO Box 601  
Cedarville, OH 45314-0601

Final Report for:  
Summer Faculty Research Program  
Wright Laboratory/Materials Directorate

Sponsored by:  
Air Force Office of Scientific Research  
Bolling AFB, Washington DC

and

Wright Laboratory/Materials Directorate

August 1995

**A REVIEW OF ELASTOMERIC SYSTEMS MADE CONDUCTIVE  
THROUGH THE USE OF CONDUCTIVE PARTICLES**

Douglas J. Miller, PhD  
Associate Professor of Chemistry  
Science & Mathematics Department  
Cedarville College

**ABSTRACT**

The literature since 1970 relating to conductive, elastomeric systems filled with carbon or metallic based particles was reviewed. Carbon based particles were normally carbon blacks, sometimes carbon fibers. Both traditional organic polymers and silicone polymers were used in these systems. The properties and use of conductive carbon blacks and metallic particles were discussed. The conductive behavior and the effects of strain or flexing on conductivity were briefly covered.

**A REVIEW OF ELASTOMERIC SYSTEMS MADE CONDUCTIVE  
THROUGH THE USE OF CONDUCTIVE PARTICLES**

Douglas J. Miller, PhD

**FOREWORD**

This article reviews the literature since about 1970 relating to conductive elastomers, especially those exhibiting conductivities over  $10^{-4}$  S/cm, which are manufactured by combining (often called filling or loading) an elastomeric substance with conductive particles. Work prior to 1970 is well summarized in a monograph.<sup>1</sup> The incredible depth and breadth of the field makes a truly comprehensive report implausible. On-line data bases, particularly Chem Abstracts, were searched for publications relating conductivity to the terms, elastomers or rubbers. Emphasis was placed on the examination of journal articles, although some patents available on-line were also obtained. Non-English language resources or patents not readily available on line were less thoroughly examined at the abstract level. Proprietary or industry related information not explicitly published in the above sources was quite possibly missed. Some elastomer manufacturers and fabricating companies were directly contacted if the literature search came up with promising leads. Material in the aforementioned monograph remains quite valuable, but much has been published since 1970, particularly on the manufacture of conductive silicone elastomers and on the use of metallic fillers.

## INTRODUCTION

Conductive elastomers, whether based on traditional organic rubbers or on silicone rubbers, have several possible uses. Those with conductivities within the range of  $10^{-12}$  to 1.0 S/cm may find use in antistatic applications.<sup>1,2</sup> Those whose conductivities exceed  $10^{-4}$  S/cm may, somewhat arbitrarily, be labeled conductive. Electrical or electronic uses, besides the control of static electricity, include wire and cable jacketing, heating elements, self-regulating resistors, electrical connectors, electronic sensors, and automotive ignition cables.<sup>1,3-5</sup> Conductive elastomers, usually in the form of gaskets, also find major use as shielding against electromagnetic interference (EMI), including radio frequency interference (RFI).<sup>4,6,7</sup> Conductive elastomers are also found in the form of coatings, caulks, sealants, and bonding agents.

## DISCUSSION: ELASTOMER TYPES

Traditional organic elastomers, or the relatively more recent silicone polymers, may be used as matrix materials in the creation of conductive elastomeric materials. Some organic rubbers frequently observed in this present literature survey were butyl rubber (isobutylene isoprene rubber), neoprene (polychloroprene), SBR rubber (copolymers of butadiene and styrene), nitrile rubber (copolymers of butadiene and acrylonitrile), EPDM rubber (terpolymers of ethylene, propylene, and a diene monomer), polyurethanes, natural rubber, and various thermoplastic compositions. Several patents, mostly foreign, mentioned the use of acrylates in films or coatings, but this present survey did extensively investigate them. Some other elastomers mentioned were fluorocarbon elastomers, polysulfides,

polythioethers, and polyepichlorohydrins. Silicone rubbers seem to have been increasingly favored over the last 20 years.

Silicone based rubbers are typically placed in a category apart from the traditional organic polymers. Their inherently different nature endues them with several advantages.<sup>8</sup> They are, therefore, frequently preferred over organic polymers for many applications.<sup>4,5,9</sup> Silicone elastomers are often more flexible than organic polymers, especially at low temperatures (down to -50 C or so), are stable over a wide range of temperatures (up into the range of 180 to 260 C), are stable in hostile environments (including oxidative), have low compression set, and are generally easy to fabricate. Fluorosilicones have the added advantages of fuel resistance and better thermal properties, but at a probable higher cost. The mechanical properties of silicone rubbers may actually be inferior at room temperature compared to other rubbers, however they change only gradually with temperature.

#### DISCUSSION: CONDUCTIVE FILLERS

The two general categories for conductive fillers are nonmetallic, usually carbon black, and metallic, usually particles of metal. Carbon filled polymers, whether organic or silicone, generally have conductivities below 1.0 S/cm with one article listing a theoretical limit of 100 S/cm based on the conductivity of pure carbon itself.<sup>10-12</sup> Carbon fibers and/or graphite may also be used alone as fillers or in combination with other conductive agents. Metallic based particles of various shapes have been utilized to reach conductivities of up to  $10^4$  S/cm.<sup>4</sup>

Carbon black is the product of incomplete oxidation or the decomposition of various fuels and has a long history of use as a reinforcing filler in organic polymers.<sup>1,10,12-22</sup> Carbon blacks can be characterized by several parameters including surface area, particle size and distribution, density, porosity, and structure. Structure has typically referred to the aggregate size and aggregate shape of carbon black particles. Void volume measurements, often by the absorption of dibutyl phthalate (DBP), have been used as a measure of structure. High structure carbon blacks have historically been blacks with a high tendency to possess aggregates of carbon particles. (It has recently been noted that the term "structure" is currently used to describe the relative void volumes of blacks with similar surface areas.<sup>12</sup>) Some carbon blacks, such as Ketjenblack EC, have high dibutyl phthalate absorption values (DBPA) due to a highly developed hollow shell type structure of each particle, and not due to aggregate characteristics (or structure).<sup>11</sup> Carbon blacks react chemically with organic rubbers during processing thus altering processing characteristics and the properties of the end product. This reinforcing effect generally results in higher tensile strength, tear resistance, and resistance to abrasive wear along with higher hysteresis and poorer dynamic performance.<sup>12</sup> An optimum range for reinforcement has been reported as 40 - 60 phr (parts by weight per 100 parts of un compounded rubber) of carbon black with a range of 35 - 80 phr listed for linear correlation of loading with hardness and modulus.<sup>23</sup> Additions of up to 200 phr of carbon black have been reported in various papers. Excessive loadings of carbon black increase the viscosity during processing to unacceptable levels and lead to

poor mechanical properties of the final product. Fillers, in general, take up between ten and fifty percent of the total volume in most fabricated elastomeric articles.<sup>21</sup> Elongation of a filled elastomer is typically decreased when filler content exceeds approximately 5% by volume. Silicone rubber has less interaction with carbon black than organic polymers, and certain mechanical properties are less affected.<sup>8,24,25</sup> Carbon blacks are used in order to produce conductive silicone elastomers, but the normal reinforcing filler for silicone rubbers is finely divided silica.<sup>8,9</sup>

Proper selection of carbon black type and processing conditions is critical for preparing conductive elastomers.<sup>10,11,13,15,16,26-28</sup> The goal is to obtain the desired level of conductivity without sacrificing good processing conditions or final mechanical properties. Carbon blacks for these purposes are selected from a group called conductive carbon blacks. Two common choices within this group are a type called acetylene blacks and a more recent brand called Ketjenblack EC, although there exists several types and brands of conductive carbon blacks. The level of carbon black necessary to impart a desired level of conductivity is much lower for conductive blacks than it is for general purpose blacks. The more recent conductive carbon blacks, such as Ketjenblack EC, Vulcan XC-72, and Philblack XE-2 have also reduced the required loading of conductive carbon black by a factor of three or four<sup>29</sup>. This level used to be 10 - 40 vol% of carbon black but now may be as low as 5 vol%.<sup>30</sup> One reported study indicated that a loading of 7 - 8% (by weight?) of a conductive black was needed to achieve a conductivity of around 1.0 S/cm, whereas 65 - 70 % of a thermal black was needed to reach the same level.<sup>12</sup> Other blacks in that

study, including acetylene black, required loadings falling within a range of 15 - 40% in order to reach conductivities near 1.0 S/cm. Highly conductive carbon blacks tend to have low density or possess aggregates of some manner (high structure).<sup>11</sup> Lower density makes it possible to place more particles of carbon black into a rubber for a given loading level, and this greater number of particles leads to higher conductivity. Aggregates tend to help in the formation of conductive pathways. Acetylene blacks have highly aggregated particle arrangements whereas Ketjenblack EC has low density due to its hollow shell geometry.<sup>11,31</sup> Ketjenblack EC does not tend to form aggregates and would normally not be called a high structure carbon black. An additional advantage to Ketjenblack EC is its lower susceptibility to mechanical damage during extensions or deformations.<sup>11</sup> Vulcan XC-72 and Philblack XE-2 also possess the hollow shell type structure.

The use of carbon fibers as conductive filler has frequently been of interest since the shape of the fibers should provide convenient pathways for conduction. There is some recent interest in such fibers alone or combined with carbon black for use as fillers in conductive elastomers.<sup>32-37</sup> Carbon fibers also exert a reinforcing influence under certain conditions, although mechanical properties decline at a certain point. Also carbon fibers generally cost more.

Metallic fillers are essentially only used to impart high conductivity to organic or silicone elastomers. They generally do not have a reinforcing effect, and several mechanical properties degrade as filler content goes up. Pure metallic particles may be used, or the metal may be coated upon another



material. The metallic particles may be in the form of fine powders, fibers, irregular flakes, or spheres. A few of the many that have been used are copper, silver, nickel, gold, steel, brass, iron, aluminum, tin, and zinc. The noble metals and nickel tend to have fewer problems involving corrosion. Sometimes one metal may be layered over another to gain an advantage in weight, cost, or conductivity. Layering a metal over an inert material, such as glass spheres, has also been done. Silver is often coated over glass, copper, nickel, or aluminum in commercially available products designed for the EMI shielding market.<sup>6,7</sup> The majority of these commercial products use silicone rubbers, although there may be significant exceptions. Conductive oxides and various alloys have been used as conductive fillers in other fabricating attempts. Elastomers filled with these metal-based particles often have conductivities within the range of 10 to  $10^4$  S/cm while other formulations yield lower values.<sup>4</sup> Silver based systems tend to have the highest conductivities (and cost).

#### DISCUSSION: CONDUCTIVITY BEHAVIOR

Conductivity of many filled elastomeric systems, whether filled with metal or carbon black, exhibit a percolation behavior. The filler material is assumed to have a very high conductivity and is assumed to be distributed throughout an insulative medium.<sup>14,38</sup> The lowest levels of loading produces no significant rise in the conductivity of the elastomer. A percolation threshold (sometimes called "critical loading" or "critical volume") is eventually reached as the loading level is increased, and the conductivity increases sharply over a relatively short range. The conductivity at the end of this range increases only gradually with additional loading. A

characteristic conductivity value vs carbon black content curve is shown in figure 1.

The percolation behavior of conductivity in carbon filled systems has often been studied and described.<sup>12,14,16</sup> Low loadings of carbon produce no conductive pathways in the rubber, and conductivity values remain near that of the unfilled polymer. Pathways, or at least "effective" pathways, begin to form at the percolation threshold. Conductivity rapidly shoots up as the number of effective pathways increase. Early workers usually envisioned a system of pathways formed by the actual contact or touching of conductive particles. More recent work favors the concept that conductivity in this region occurs from the movement of electrons across the gaps between particles or between aggregates of particles.<sup>11,12,14,28,39, 40</sup> Increasing the number of particles or aggregates causes the size of the gaps to decrease, and conductivity rapidly increases in this percolation region. Some mechanisms or terms used to describe electron movement across the gaps are jumping, tunneling, hopping, dielectric breakdown, and Schottky conduction.<sup>11,14,28,39-41</sup> Additional loading eventually causes actual contact of the particles.<sup>14, 39,40</sup> Increasing the loading at this stage packs the particles more tightly together thus reducing the contact resistance. Once contact resistance is minimized, further loading would not significantly affect the conductivity. Loadings of over 30 wt% of acetylene black in silicone rubber reportedly resulted in particle contact and conductivities of 0.1 S/cm or higher.<sup>39,40</sup> Acetylene black loadings in the range of 15 - 30 wt% reportedly gave rise to Schottky conduction across the gaps between aggregates. Conductivity within this range rapidly changed as

the gap distances changed. Systems with less than 15 wt% of carbon black were said to exhibit insulative behavior.

The conductivity of metal-filled elastomeric systems has not been as thoroughly explored in the journal literature as have the carbon black filled systems. Reports on metal filled polymers primarily discuss conductive plastics, but sometimes include elastomers.<sup>42,43</sup> Much of the discussion in these reports is applicable to elastomers. Percolation behavior of the conduction in metal filled polymers is observed, but the transition from insulator to conductor is more abrupt than in carbon black rubbers.<sup>22</sup> Most workers assume that conduction after the narrow percolation region occurs mainly due to particle contact.<sup>22,30,44,45</sup> References 42 and<sup>44</sup> contain good discussions about the conductivity in metal filled materials. One of these references is an excellent and comprehensive study on certain commercially available, metal filled, silicone rubbers.<sup>44</sup> It is typical to report the critical volume,  $V_c$ , required to establish electrical continuity in these systems. The reported values of  $V_c$  varies widely but are commonly within the range of 10 to 50 vol%. The study on commercial, silicone rubbers had filler content (Ni, Au-plated Ni, Ag-plated Ni, or Ag) which varied from 25.7 to 39.5 vol%. This translated into 78.3 to 85.2 wt%. Another study on the critical volumes in silicone rubber reported values of 5 to 20 vol% for powdered silver filler and 15 to 50 vol% for silver coated spheres.<sup>45</sup> This last study also reported that  $V_c$  decreased when smaller particles of fine powders were used or when fibers or flakes with an aspect ratio greater than one were used.  $V_c$  increased when spherical powders were used in place of fine powders. Flakes and fibers will normally become

conductive at lower loadings than irregular particles of low aspect ratios.<sup>46,47</sup> A 1 vol% concentration of high aspect stainless steel fibers in a polymer, probably a plastic, had a conductivity of 1.4 S/cm.<sup>47</sup> High aspect ratio stainless steel fibers in silicone rubber at concentrations of under 3.0 vol% provided good shielding effectiveness (SE) against EMI.<sup>46</sup>

The above-mentioned study on commercial, conductive, silicone rubbers discussed the various factors affecting conductivity and may contain elements applicable to organic elastomers.<sup>44</sup> Silver plating of nickel particles did not improve the conductivity over uncoated nickel particles. Filling silicone rubbers with metallic particles in order to achieve a certain conductivity can increase the polymer's modulus to undesirable levels. Lowering the required metal loading to reach a given conductivity or increasing the conductivity at a given loading level is a desirable goal. One way to accomplish this is to choose a metal with higher conductivity. An ordering of conductivities for some of the commonly used metals is, from highest to lowest: silver, copper, gold, aluminum, brass, nickel, iron, tin, and steel.<sup>48</sup> Steel and iron are magnetic. The others are nonmagnetic. The largest contribution to resistance in metal filled systems is the interparticle microresistance (constriction resistance).<sup>22,30,44,45</sup> The more deformation that there is where two particles meet, the lower the constriction resistance. Softer metals would, therefore, lead to higher conductivities. Annealed copper, aluminum, and tin are softer than silver or nickel. Particles with higher aspect ratios, such as fibers, would be expected to yield higher conductivity. Processing conditions can also affect conductivity. Overprocessing usually leads to lower conductivity.

## DISCUSSION: EFFECTS OF STRAIN OR FLEXING

The effect of normal strain upon most filled, conductive elastomers is a lowering of the conductivity values. The 1970 monograph was the best source for general information on the effects of deformations on organic rubbers.<sup>1</sup> It is somewhat surprising that no recent studies with overall implications in this field were located. This may be due to the complexity seen in this field, the lack of reproducibility, and to the fact that each conductive polymer configuration has its own unique characteristics. The earliest studied examples were, of course, carbon filled organic polymers.<sup>1,13,14,16</sup> Large strains over 30% tend to result in complex conductivity changes depending on the exact conditions. An older study on the effects of cyclic stretching of conductive, carbon black filled rubbers gave a ranking (from worst to best) of nitrile rubber, SBR, EPDM, natural rubber, polychloroprene, and butyl rubber.<sup>16</sup> The effect of strain is also more evident in rubbers with conductivities falling within the percolation region in which a slight change in the effective concentration of carbon black may cause a large change in conductivity. The percolation region for most carbon filled systems usually occurs below a conductivity of  $10^{-4}$  S/cm. Another older study has been reported to claim that conductive silicones were relatively insensitive to strain.<sup>1,16</sup> Recovery of conductivity after small elongations (less than 30%) may be nearly complete if the elastomer is given time to recover. Mild heating may accelerate this effect.

Recent sources (located during this survey) of general information concerning the effects of strains or distortions on conductive elastomers were somewhat rare and usually concerned with silicone rubbers. One 1976

text about carbon black did briefly discuss this subject for organic rubbers.<sup>19</sup> Elongation past 10% of commercial, conductive, particle filled silicone gaskets is not recommended due to low tensile strength.<sup>49</sup> Systems using metal coated glass spheres have been reported to be sensitive to processing conditions before vulcanization and to vibration afterwards.<sup>6,22</sup> Electromagnetic pulse (EMP) induced current may damage metal coated, glass spheres. The report referenced earlier contains a fair amount of material concerning the effects of strain upon the conductivity of four commercial, metal filled, silicone elastomers.<sup>44</sup> Conductivity was studied at different temperatures for strains under 50%, and a simple model was developed. This model, which was not followed exactly, predicted decreasing conductivity with increasing strain and more sensitivity of conductivity to strain for lower metal volume fractions.

A recent series of papers examined the use of short carbon fibers (SCF) in nitrile rubber.<sup>34-37</sup> Composites filled only with SCF were found to have percolation thresholds at lower loadings than for carbon black composites, but the SCF was determined to interact less with the rubber. Thus the composites containing only SCF showed a distinct tendency to yield in their stress-strain plots, whereas the composites filled only with carbon black did not. One of these studies compared the effects of extensional strain on the conductivity of composites with only SCF (40 to 60 phr) to those containing filler blends of SCF and carbon black. All of the composites studied had conductivities ranging from about 0.46 to 3.45 S/cm. Increasing the concentration of SCF as the sole filler resulted in less sensitivity of conductivity to strain. Elongation was said to both create conductive

pathways as a result of fiber orientation and cause breakage of pathways due to lengthening the gaps between fibers. Pathway creation almost balanced breakage at low strains (below about 20%), then breakage began to predominate. Using a blend of conductive carbon black with SCF as conductive filler while keeping the total carbon loading (fiber plus black) constant at 60 phr, resulted in a more reinforced material (less tendency to yield) with only moderately lower conductivity. The addition of carbon black helped bridge the gaps between fibers, thus making the composite's conductivity less sensitive to strain. This addition of carbon black also improved other characteristics compared to fiber only composites. Composites with only SCF as a filler showed significant tensile set in their stress-strain plot after elongation, and their conductivity after extension was substantially higher than originally measured. A composite filled with SCF (40 phr) and carbon black (20 phr) exhibited these effects, especially the conductivity change, to a smaller degree than a similar system filled only with 60 phr SCF.

## REFERENCES

- (1) Norman, R. H. *Conductive Rubbers and Plastics*; Elsevier Publishing Company Limited: New York, 1970.
- (2) Harding, J. A.; Gerteisen, S. R. *Proc., Annu. Tech. Conf. - Soc. Plast. Eng.*, 49th 1991, 2409-13.
- (3) Janson, G. *Mater. Des.* 1991, 12(3), 133-137.
- (4) Kroupa, L. *Rubber World* 1989, 200(3), 23-4, 26-8.
- (5) Piccirillo, T. P.; Dalamangas, C. A.; Buchoff, L. S.; Hasan, R. *Proc., Holm Seminar on Electrical Contacts* 1976, 71-78.
- (6) EMI Shielding Engineering Handbook; Chomerics.
- (7) EMI Shielding Products; TECKNIT.
- (8) Hardman, B. B.; Torkelson, A. In *Encyclopedia of Chemical Technology*, 3rd ed.; Grayson, M.; Eckroth, D.; Mark, H. F.; Othmer, D. F.; Overberger, C. G.; Seaborg, G. T., Eds.; John Wiley & Sons, Inc.: New York, 1982; Vol. 20, pp. 922-962.
- (9) Wolfer, D. *Eur. Rubber J.* 1977, 159(4), 16, 18-19, 22-3.
- (10) Van Drumpt, J. D. *Plast. Compd.* 1988, 11(2), 37, 40, 42, 44.
- (11) Verhelst, W. F.; Wolthuis, K. G.; Voet, A.; Ehrburger, P.; Donnet, J. B. *Rubber Chem. Technol.* 1977, 50(4), 735-746.
- (12) Dannenberg, E. M.; Paquin, L.; Gwinnell, H. In *Encyclopedia of Chemical Technology*, 4th ed.; Howe-Grant, M., Ed.; John Wiley & Sons, Inc.: New York, 1992; Vol. 4, pp. 1037-1074.
- (13) Brokenbrow, B. E.; Sims, D.; Stokoe, A. L. *Rubber J.* 1969, 151(12), 30-1, 33, 36, 38, 40, 42, 44, 46, 49, 51.
- (14) Medalia, A. I. *Rubber Chem. Technol.* 1986, 59(3), 432-54.
- (15) Juengel, R. R. *Rubber World* 1985, 192(6), 30-35.
- (16) Pyne, J. R. *Eur. Rubber J. Urethanes Today* 1981, 163(9), 17, 19-20.
- (17) Thayer, A. M. *Chemical & Engineering News* 1995, 73(29), 33, 35, 37-40.
- (18) Medalia, A. I. In *Carbon Black-Polymer Composites*; *Plastics Engineering 3*; Sichel, E. K., Ed.; Marcel Dekker, Inc.: New York, 1982; pp. 1-49.



- (19) Donnet, Jean-B.; Voet, A. *Carbon Black*; Marcel Dekker, Inc.: New York, 1976.
- (20) Sichel, E. K. *Carbon Black-Polymer Composites*; *Plastics Engineering* 3; Marcel Dekker, Inc.: New York, 1982.
- (21) Falcone, J. S., Jr. In *Encyclopedia of Chemical Technology*, 4th ed.; Howe-Grant, M., Ed.; John Wiley & Sons, Inc.: New York, 1993; Vol. 4, pp. 745-761.
- (22) Reboul, J. P. In *Metal-Filled Polymers*; *Plastics Engineering* 11; Bhattacharya, S. K., Ed.; Marcel Dekker, Inc.: New York, 1986; Chapter 6, pp. 335-354.
- (23) Dannenberg, E. M. In *Encyclopedia of Chemical Technology*, 3rd ed.; Grayson, M.; Eckroth, D.; Mark, H. F.; Othmer, D. F.; Overberger, C. G.; Seaborg, G. T., Eds.; John Wiley & Sons, Inc.: New York, 1978; Vol. 4, pp. 631-666.
- (24) Pouchelon, A.; Vondracek, P. *Rubber Chem. Technol.* 1989, 62(5), 788-799.
- (25) Morton, M. In *Encyclopedia of Polymer Science and Technology*, 4th ed.; Howe-Grant, M., Ed.; John Wiley & Sons, Inc.: New York, 1993; Vol. 4, pp. 905-923.
- (26) Swor, R. A.; Harris, D. R.; Lyon, F. *Kautsch. Gummi, Kunstst.* 1984, 37(3), 198-206.
- (27) Sircar, A. K.; Lamond, T. G. *Rubber Chem. Technol.* 1978, 51(1), 126-132.
- (28) Polley, M. H.; Boonstra, B. B. S. T. *Rubber Chem. Technol.* 1957, 30, 170-179.
- (29) Geuskens, G.; Gielens, J. L.; Geshef, D.; Deltour, R. *Eur. Polym. J.* 1987, 23(12), 993-995.
- (30) Kusy, R. P. In *Metal-Filled Polymers*; *Plastics Engineering* 11; Bhattacharya, S. K., Ed.; Marcel Dekker, Inc.: New York, 1986; Chapter 1, pp. 1-142.
- (31) Smith, W. R.; Bean, D. C. In *Encyclopedia of Chemical Technology*, 2nd ed.; John Wiley & Sons, Inc.: New York, 1964; Vol. 4, pp. 278-280.
- (32) Ruckenstein, E.; Hong, L. J. *Appl. Polym. Sci.* 1994, 53(7), 923-932.
- (33) Jana, P. B.; De, S. K.; Chaudhuri, S.; Pal, A. K. *Rubber Chem. Technol.* 1991, 65, 7-23.

- (34) Pramanik, P. K.; Khastagir, D.; Saha, T. N. *J. Mater. Sci.* 1993, 28(13), 3539-3546.
- (35) Pramanik, P. K.; Khastgir, D.; Saha, T. N. *Plast., Rubber Compos. Process. Appl.* 1992, 17(3), 179-185.
- (36) Pramanik, P. K.; Khastgir, D.; Saha, T. N. *Composites* 1992, 23(3), 183-191.
- (37) Pramanik, P. K.; Khastgir, D.; Saha, T. N. *J. Elastomers Plast.* 1991, 23(4), 345-361.
- (38) Kirkpatrick, S. *Rev. Mod. Phys.* 1973, 45(4), 574-588.
- (39) Tamai, T. *IEEE Trans. Compon., Hybrids, Manuf. Technol.* 1982, Chmt-5(1), 56-61.
- (40) Tamai, T. In *Physicochem. Aspects Polym. Surf., [Proc. Int. Symp.]*, Meeting Date 1981, Volume 1; Mittal, K. L., Ed.; Plenum: New York, N. Y., 1983; pp. 507-520.
- (41) Sichel, E. K.; Gittleman, J. I.; Sheng, P. In *Carbon Black-Polymer Composites*; *Plastics Engineering 3*; Sichel, E. K., Ed.; Marcel Dekker, Inc.: New York, 1982; pp. 51-78.
- (42) Bhattacharya, S. K., Ed. *Metal-Filled Polymers*; *Plastics Engineering 11*; Marcel Dekker, Inc.: New York, 1986.
- (43) Delmonte, J. In *Metal/Polymer Composites*; Van Nostrand Reinhold: New York, 1990; Chapter 4, pp. 77-101.
- (44) Pike, G. E. *Electrical Properties of Conducting Elastomers*; Report, SAND-81-0263, 78 pp. Avail. NTIS From: Energy Res. Abstr. 1981, 6(14), Abstr. No. 20522.
- (45) Ruschau, G. R.; Newnham, R. E. *J. Compos. Mater.* 1992, 26(18), 2727-2735.
- (46) Bigg, D. M. In *Metal-Filled Polymers*; *Plastics Engineering 11*; Bhattacharya, S. K., Ed.; Marcel Dekker, Inc.: New York, 1986; Chapter 3, pp. 165-226.
- (47) Bigg, D. M.; Bradbury, E. J. In *Conductive Polymers*; *Polymer Science and Technology 15*; Seymour, R. B., Ed.; Plenum Press: New York, 1981; pp. 23-38.
- (48) Design Guide to the Selection and Application of EMI Shielding Materials; TECKNIT.

- (49) Conductive Elastomer Gasket Design; In EMI Shielding Engineering Handbook; Chomerics.

Any tables and figures referred to in this version of the final report sent to the Air Force Office of Scientific Research have been omitted due to the page limit constraints. This material plus an expanded report, including a section on specific systems and examples, may be obtained by contacting:

WL/MLBP  
2941 P STREET -- SUITE 1  
WPAFB, OH 45433-7750

# A SYSTEMS STUDY OF THE SCALABLE COHERENT INTERFACE (SCI)

Sarit Mukherjee  
Assistant Professor  
Department of Computer Science & Engineering

University of Nebraska-Lincoln  
115 Ferguson Hall  
Lincoln, NE 68588-0115

Final Report for:  
Summer Faculty Research Program  
Wright Laboratory

Sponsored by:  
Air Force Office of Scientific Research  
Bolling Air Force Base, DC

and

Wright Laboratory

August 1995

# A SYSTEMS STUDY OF THE SCALABLE COHERENT INTERFACE (SCI)

Sarit Mukherjee  
Assistant Professor  
Department of Computer Science & Engineering  
University of Nebraska-Lincoln

## Abstract

The SCI is a new high-speed multiprocessor interconnection standard that delivers GBytes/sec transmission rate along unidirectional point-to-point links connected into a ring topology. Because of the cost and performance potential offered by SCI, the Joint Advanced Strike Technology (JAST) program has selected SCI and its unspecified derivative SCI Real-Time (SCI/RT), as the baseline architecture to address the needs of military aircraft in the post-2005 time frame. The performance of SCI in supporting different classes of applications has not been studied extensively. A need was felt to test different avionics applications on SCI before the actual chip-set is manufactured. The use of extensive simulation of the SCI protocol and evaluation of its performance, when subjected to avionics applications is the means towards this end. This report outlines the work conducted on SCI at the Wright Laboratory during the summer of 1995. As a part of the Summer Faculty Research Program of the Air Force Office of Scientific Research (AFOSR), the author designed and implemented a simulator for the SCI ring at the Wright Laboratory at the Wright Patterson Air Force Base. The kernel of the simulator is the basic SCI transaction model. The simulator, built on the kernel using C programming language in an Unix environment, has the capability of supporting many of the SCI features and interconnection topologies. This report describes the work conducted and the feasibility study of SCI.

# A SYSTEMS STUDY OF THE SCALABLE COHERENT INTERFACE (SCI)

Sarit Mukherjee

## 1 Introduction

Large scale distributed memory processor networks or massively parallel processors (MPP) have become the computers of choice for large computationally intensive tasks in recent years. MPP architectures consist of a set of nodes where nodes consist of processors(s), local memory, message router, and other support devices. MPP architectures often connect nodes through direct network in which each node has a connection to a set of other nodes, called neighbors. Since memory is distributed, MPP nodes communicate by sending messages through the network. The Scalable Coherent Interface (SCI) is standardized for very high performance multiprocessor systems that supports a coherent shared memory model scalable to systems with up to 64K nodes.

Because of the cost and performance potential offered by the SCI concept, the Joint Advanced Strike Technology (JAST) program has selected SCI and its unspecified derivative SCI Real-Time (SCI/RT) as the baseline architecture to address the needs of military aircraft in the post-2005 time frame. JAST requirements were defined by several system studies, including Air Force PAVE PACE efforts and the Navy's Next Generation Computer Resource Program. SCI/RT is intended to be an enhancement for SCI which improves the real time and fault tolerance capabilities of SCI. Currently, the Air Force and Navy are jointly involved in two separate contracts in which SCI-based hardware is being developed.

Not a whole lot of work has gone into the performance study of SCI [6, 2]. A need was felt to test different avionics applications on SCI before the chip-set is manufactured. The clear and viable choice is extensive simulation of SCI protocol and evaluation of its performance, when subjected to avionics applications. This report outlines the work conducted by the author on SCI at the Wright Laboratory during the summer of 1995. As a part of the the Summer Faculty Research Program of Air Force Office of Scientific Research (AFOSR), the author designed and implemented a simulator for the SCI ring at the Wright Laboratory at the Wright Patterson Air Force Base. The kernel of the simulator is the basic SCI transaction model. The simulator, built on the kernel using C programming language in an Unix environment, has the capability to support many of the SCI features and interconnected topologies. This report describes the simulation and the feasibility study of SCI.

The rest of the document is organized as follows: In section 2 we present an overview of the SCI node architecture, protocol and SCI switch. Sections 3 and 4 describe the simulation model and the experimentation. The report is concluded in section 5 with mention to some future directions.

The source code of the simulator is presented in the appendix.

## 2 Overview of SCI

The SCI is a new high-speed multiprocessor interconnection standard [4] that delivers GBytes/sec transmission rate along unidirectional point-to-point links that is connected into a ring topology. SCI was developed by a working group of leading computer researchers who wished to overcome the fundamental physical limits imposed by bus technology. SCI provides the services of a backplane. Figure 1 shows the basic topology of a SCI ring.

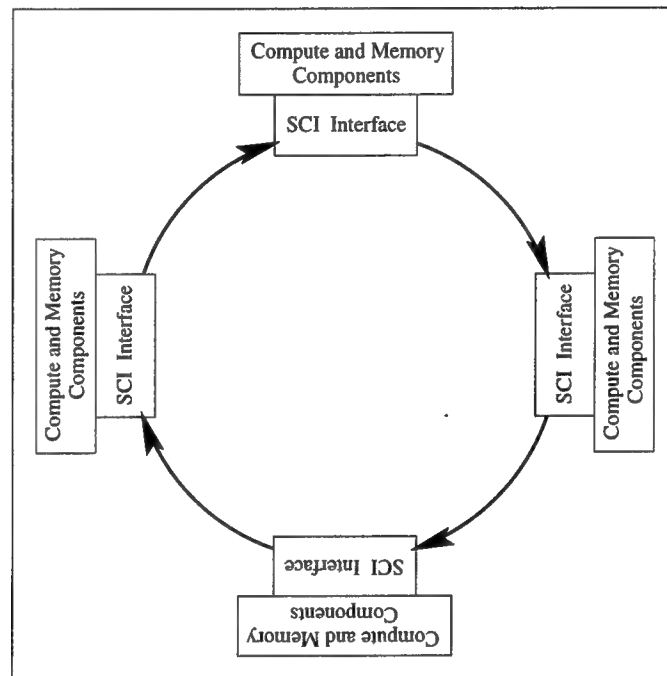


Figure 1: The basic SCI ring.

The SCI interface (also referred to as SCI node) is the unit through which the compute and memory components communicate with other compute and memory components connected to the ring. Its logical queueing structure is identical to that of a buffer-insertion ring interface [1] (see figure 2). The node interface consists of two unidirectional links (input and output) which are used to connect nodes in a ring topology. The bypass FIFO stores packets arriving from upstream neighbor while the node is transmitting packets. This enables a node to concurrently (1) transmit packets, (2) process packets addressed to other nodes, and (3) accept packets addressed to itself.

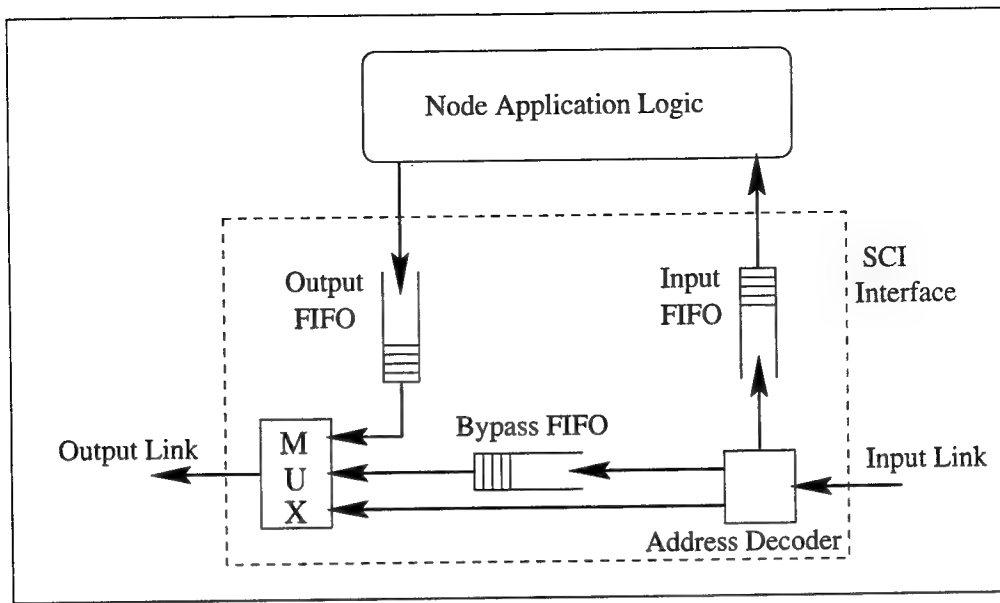


Figure 2: SCI interface (also referred to as SCI node).

## 2.1 Basic SCI Protocol

The steps executed in the basic SCI protocol are as follows [6]:

- A node always transmits a symbol onto its output link. If a node has no packet to send, it transmits an idle symbol.
- When a source node desires to send a packet
  - It places the packet in its output FIFO.
  - If the bypass FIFO is empty, and the node is not currently transmitting a packet from address decoder, the send packet is immediately output onto the ring.
  - A copy of the packet is saved at the source.
- Upon arrival at the downstream node
  - A send packet is parsed (also known as address decoding).
  - It is either stripped (passed to the input FIFO), or passed along the ring.
  - In absence of contention, a passing packet may directly be routed to the output link.
  - If the bypass FIFO is not empty, or the output FIFO is currently transmitting, the packet is routed to the bypass FIFO.



- When the output FIFO completes transmission
  - If bypass FIFO has accumulated symbols, output resumes from the bypass FIFO. This is called the recovery stage.
  - The node is not allowed to transmit another packet during the recovery stage.
- When the packet reaches its target node
  - It is stripped and placed into the input FIFO or discarded if the input FIFO is full.
  - The target node sends an echo packet to the source.
- When the echo packet reaches the source node
  - It is matched with the saved copy of the send packet.
  - The saved packet is discarded if the transmission was successful.
  - The saved packet is retransmitted if the transmission was unsuccessful.

## 2.2 SCI Transactions

The SCI transaction model [4] is shown in figure 3. A transaction can be broken into request

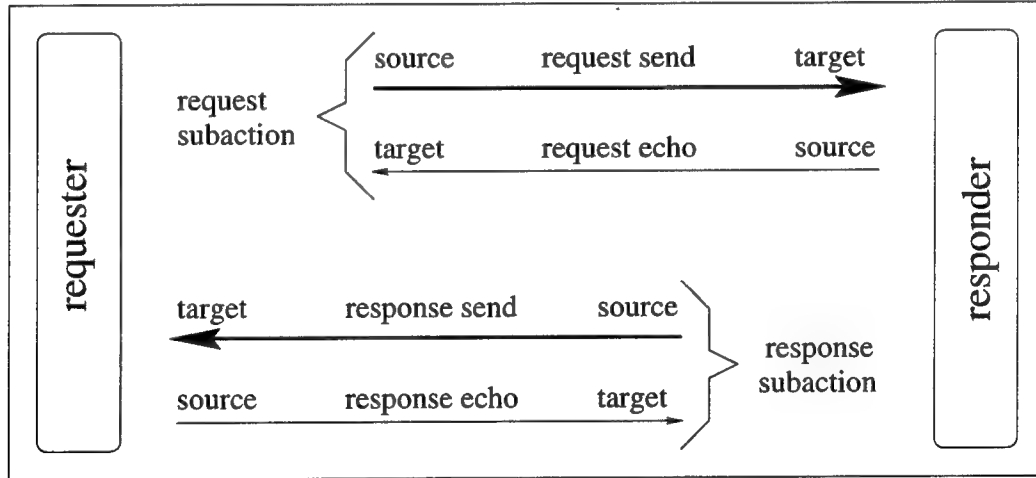


Figure 3: SCI transaction model.

sub-action and response sub-action. A sub-action consists of the source node transmitting a send packet to the target node. Upon receipt of a send packet, the target node returns an echo packet to the source. The echo packet indicates whether the target node accepted the send packet. If the

returned echo packet indicates that the send packet is not accepted, the source node retransmits the packet.

There are different types of transactions specified in SCI, which are broadly divided into two classes: response expected (e.g., read, write) and responseless (e.g., move). While the former requires both request and response sub-actions, the latter requires only the request sub-action.

## 2.3 SCI Requester-Responder Model

As the SCI transaction model suggests, a node in a SCI ring could be a requester, or a responder or both. In our model we consider the most general full-duplex nodes having the capability of serving as a requester as well as a responder. In order to avoid system deadlock in such a node, request and response sub-actions are processed through separate queues [4], as shown in figure 4. Only one bypass buffer is introduced for cost and performance purpose. In order to take care of speed mismatch between the SCI interface and the node application logic, queues are introduced in the SCI interface (see figure 4).

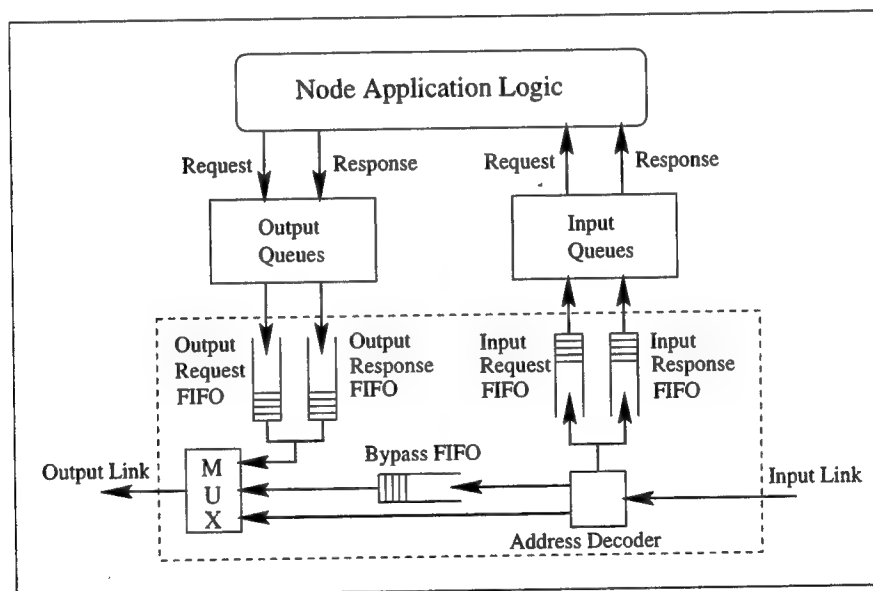


Figure 4: SCI node with dual requester and responder capability.

## 2.4 SCI Switch

The standard defines a switch in the following way (page 45 of [4]): *A device that connects ringlets and has queues. It may behave as a consumer (when accepting remote transactions) and as a*

producer (when forwarding the sub-action to another ringlet). It may be visible as a node with a node identifier, or be transparent with no node identifier.

This definition was the starting point of our switch design. Note that the SCI standard does not strictly define the internal architecture of a switch. The design is left to the vendor as long as the specified switch functions are provided. As we will describe in section 3, our switch model is very modular and parametric so that different switching options can be simulated and tested very easily.

The basic model of switch is shown in figure 5. It connects two rings (rings X and Y). The

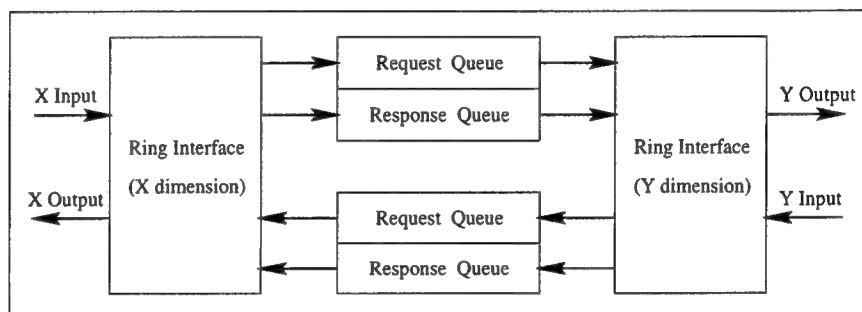


Figure 5: Basic SCI switch.

queues are provided to take care of the speed mismatch, if any, between the two rings. Although not shown in the figure, the dimension of a switch can be more than two, i.e., a switch can connect more than two rings.

SCI switch is the basic component used to extend the SCI rings to other more complicated interconnected topologies. As we will show in section 4, the topology of the interconnected rings plays a very important role in the system performance. Some of the topologies are shown in figure 6.

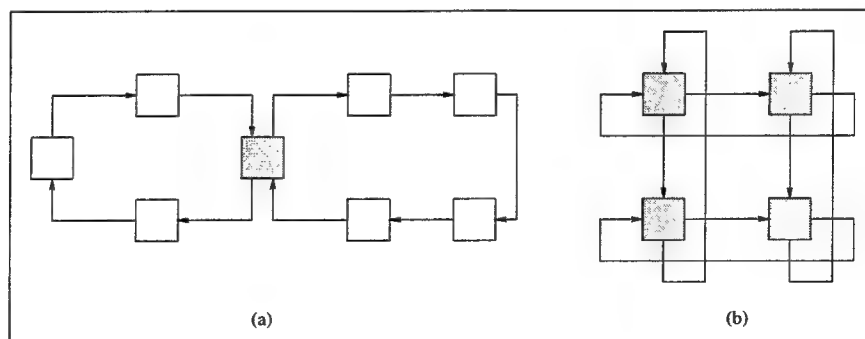


Figure 6: Different SCI topologies: (a) Two Interconnected Rings, and (b) 2-dimensional Torus. Shaded nodes are switches.

The SCI switch can be classified into two different classes:

**Agent:** In an agent switch the send packet is forwarded through the switch towards the destination.

The switch generates and sends echo packet back to the source just as if it were the target node of the send packet.

**Agentless:** In an agentless switch the send packet is forwarded through the switch towards the destination. The switch forwards, *but does not generate* echoes between the source and the target.

### 3 Simulation Model of SCI

In this section we detail the modeling of different components that were used in the simulation of SCI. The simulator was developed using C programming language in a Unix environment. We started with a baseline simulation model developed at the University of Wisconsin. Below we outline the features and shortcomings of the Wisconsin simulator.

#### 3.1 The Baseline Simulator

The simulator developed at the University of Wisconsin [6] was used as the baseline simulator for our purpose. The features and shortcomings of the baseline simulator are as follows:

1. It simulates a single ring with variable number of SCI nodes.
2. No cache coherence protocol of SCI was simulated.
3. The simulator does not differentiate between the requester and the responder. Therefore, it cannot model different kinds of SCI transactions.
4. It uses simplified buffer management — single transmit and receive queue per SCI node.
5. It does not model the node application logic, e.g., CPU, memory, cache, nor does it model the transfer module (Processor Interface Unit) responsible for communicating between the SCI interface and the node application logic (to be explained shortly).

#### 3.2 Simulation Model of the Overall System Node

We made several extensions to the baseline model to make it more conformant to the SCI standard and useful for our feasibility study. In the following, we briefly outline the main extensions, and then discuss some of the aspects in more detail.

1. We designed and implemented System Node and Interface Model. It includes the dual requester and responder model as shown in figure 4, application node model (e.g., CPU, memory, cache) and the transfer module (also known as processor interface unit) as shown in figure 7.

2. We implemented several SCI transactions, belonging to response expected as well as responseless classes.
3. We designed and implemented the SCI switch. This was used to simulate several interconnected topologies.

The figure 7 shows the overall system node model that was simulated. The compute unit is

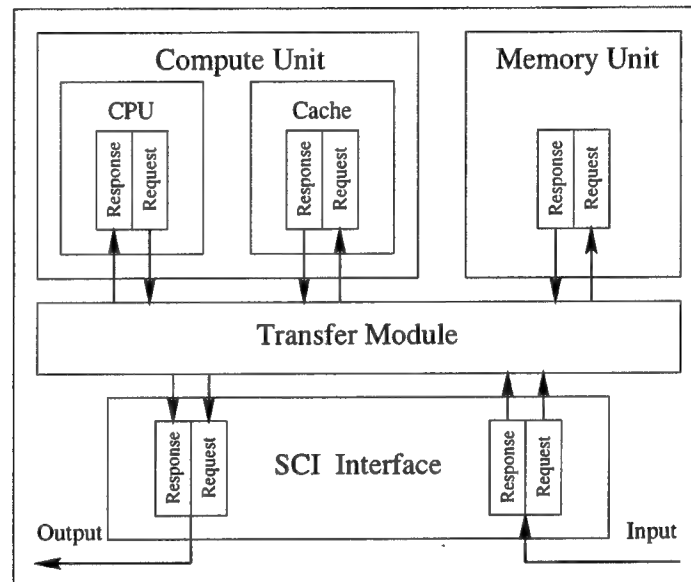


Figure 7: System node model used in simulation.

composed of CPU and cache (we can also distinguish level 1 and level 2 cache in our model). The CPU is responsible for simulating the task execution. The memory unit is the system memory component. In our model, the memory controller can work concurrently with the CPU to respond to external requests. The SCI interface models the dual mode SCI node as described in section 2 and the simulated node model is exactly the same as shown in figure 4. The transfer module is responsible for the bidirectional communication between the SCI interface and the compute and memory units. In order to take care of speed mismatch among the different units and to support concurrency, queues are introduced in the node model.

When a task requires remote operations (e.g., fetching some data which is not in its local memory, writing into a remote cache, etc.), the remote request is entered into the CPU's request queue (note that CPU is the only component that can generate requests). The transfer module is responsible for transferring the request to the SCI's outgoing request queue. When a response to a request is generated (note that response is generated by cache or the memory unit), it follows the reverse path and eventually gets entered into the response queue of the CPU.

In our simulation model, the speed and bandwidth of different components can be varied independently. This gives the user a lot of flexibility in simulating different system architectures.

### 3.3 Simulation Model of Switch

The simulated switch model is shown in figure 8. It is composed of two SCI interfaces placed

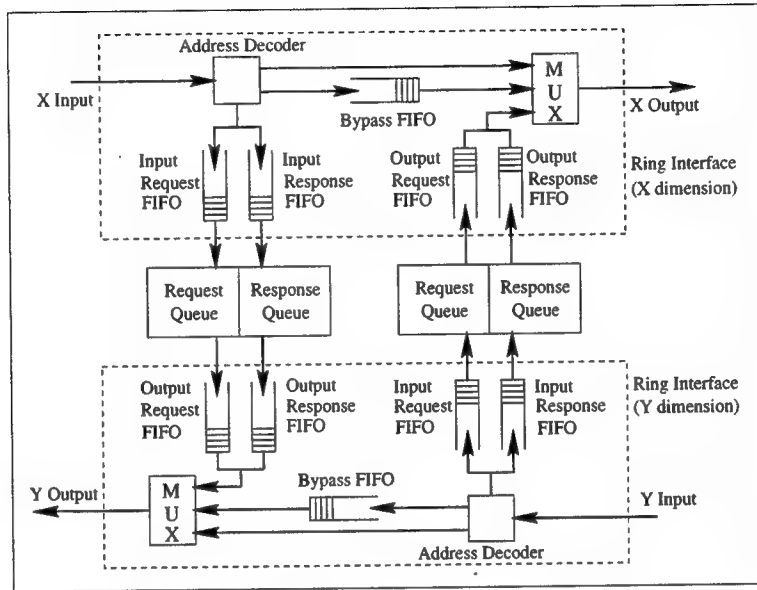


Figure 8: Simulation model of the SCI switch.

back-to-back. Each interface belongs to a different ring. Again queues are introduced to take care of speed mismatch between the rings. Although it is not shown in the diagram, the switch is modeled in a fashion that makes multi-dimensional switch (dimension more the two) simulation a very easy extension. Moreover, the switch design is modular. This allows to change different components of the switch very easily. For example, the routing algorithm in the switch is different for different topologies. In our model, introduction of a new routing scheme in the switch amounts to just changing a parameter in a function call. The switch can be either a passive or an active unit. In a passive mode, the switch can only exchange packets between the rings, but cannot have its own node application logic. The active mode of operation allows the switch to have compute and memory components (i.e., the switch itself can generate and receive packets). The switch can determine the shortest route to the destination from the alternative paths.

### 3.4 Simulation Software Structure

In this section we present an overview of the structure of the simulator and its different components. It is a time-driven simulator. The clock of the fastest component<sup>1</sup> was used to define the unit of time. The reason behind using the time-driven approach is to simulate minute details of the SCI protocol and the node application logic very elaborately and to compute different performance metric very accurately. The simulator is very modular in design and is organized in a hierarchical fashion. In the rest we present an overview of the structure of the simulator. The details can be found in the appendix.

There is a master module in the simulation which invokes other modules within each node connected in the interconnected system. It can simulate different speed of each component very easily, as long as unit of time is determined from the fastest clock. Moreover, note that it calls only the `SCI_Node()` module to simulate both the simple SCI node and the switch. This is achieved by making a node intelligent enough to work as a switch as well as a simple node. This design allows easy extension to multi-dimensional switch.

#### Master Module

```
Initialize the system
time = 0
while (time < END_TIME) do begin
    for each node of the system do
        if (time MOD cpu_clock_ratio == 0)
            CPU_Node()
        if (time MOD cache_clock_ratio == 0)
            Cache_Node()
        if (time MOD memory_clock_ratio == 0)
            Memory_Node()
        if (time MOD transfer_clock_ratio == 0)
            Transfer_Module()
        if (time MOD sci_clock_ratio == 0)
            SCI_Node()
    end
    time = time + 1
end
```

---

<sup>1</sup>For most cases we studied it was the SCI clock.

Below we outline the functionalities of each called module. The CPU\_Node() is responsible for processing tasks and remote request generation. The initial state of the CPU could be COMPUTE, SUSPEND or REQUEST depending on the task it is executing. This model allows flexibility and also simulates response expected and responseless transactions of SCI very efficiently.

```

CPU_Node()

case task_state of begin
  REQUEST:
    Prepare the Request Packet
    Enter the Request Packet in the Request Queue
    task_state = SUSPEND
  RESPONSE:
    Receive the Response Packet from the Response Queue
    task_state = COMPUTE
  COMPUTE:
    Perform computation
    If remote operation needed
      task_state = REQUEST
  SUSPEND:
    Check Response Queue for packet arrival
    If packet has arrived
      task_state = RESPONSE
end
```

There is a single software module that can simulate the cache unit as well as the memory unit. By differing the frequency of calls to this module, it can perform the job of a fast cache component or a relatively slow memory component. This module starts from CHECK state.



### **Cache\_Node() or Memory\_Node()**

```
case state of begin
  RECEIVE:
    Receive the Request Packet from the Request Queue
    Perform packet transaction (e.g., Read, Write)
    state = TRANSMIT
  TRANSMIT:
    Prepare the Response Packet
    Enter the Response Packet in the Response Queue
    state = CHECK
  CHECK:
    Check Request Queue for packet arrival
    If packet has arrived
      state = RECEIVE
end
```

The next module describes the functions of the transfer module. Although not mentioned in the description, the queues holding requests and responses may actually be residing in the local memory (e.g., the request and response queues of the CPU should actually be in the memory). This module is designed to take care of the speed mismatch in access time depending on the location of the queues.

### **Transfer\_Module()**

```
Poll incoming queues
If there is a packet present in the polled queue
  Receive the packet
  Enter the packet in the appropriate outgoing queue
```

We designed one software module to serve the purpose of both a simple SCI interface (or node) and a more complicated SCI switch. The routing mechanism implemented at the module can detect whether to accept a packet in the simple node mode or accept and forward the packet in the switch mode. Moreover, by changing the routing function, we can very easily simulate different interconnected topologies.

### SCI\_Node()

Forward a symbol to downstream node using the output link;  
Accept packets from the node application logic (the switch  
buffer, if the node is a switch) to output FIFO;  
Deliver packets routed to this node from input FIFO to the  
node application logic (the switch buffer, if the node is a switch);  
Process packets coming from upstream node on the input link;

## 4 Simulation Experiments

The purpose of our simulation experimentation was (1) to verify the design and implementation of the complete system, and (2) to study the impact of topology (i.e., switch) on the performance of the interconnected system. In particular we were interested in determining the impact of flow control and traffic locality. Two extreme traffic locality scenarios were simulated using uniform traffic and hot-spot traffic. In the uniform traffic case, message traffic generation is equally likely at each node in the system, and message traffic destination is equally likely to be any particular node in the system. In the hot spot traffic pattern, one node attempts to use as much ring bandwidth as possible, and the message traffic destinations are equally likely. The performance metrics we compared were throughput (measures number of bytes received per unit time) versus latency (measures the time taken by a message to get from the source to the target). The throughput-latency curves indicate how the system performance is affected when traffic load is varied.

Figure 9 shows the different topologies simulated and studied. The shaded box denotes a switch and the white box is a simple SCI node. Note that all the interconnected systems have the same number of sixteen nodes. The simulation parameters are shown in the following table.

Component	Clock (MHz)	Width (bytes)	Other
SCI	500	2	Internode clock cycle = 4 60% address packets, 40% data packets
CPU	100	8	
Cache	100	64	
Memory	20	64	
Transfer Module	500	64	

We have run several experiments on the simulated topology. The parameters that were varied are the traffic pattern (uniform or hot-spot traffic) and flow control of SCI (with or without active flow control). In each experiment we measured the throughput-vs-latency to determine the performance of each of the interconnected systems. The results from the runs are plotted in figures 10 and 11.

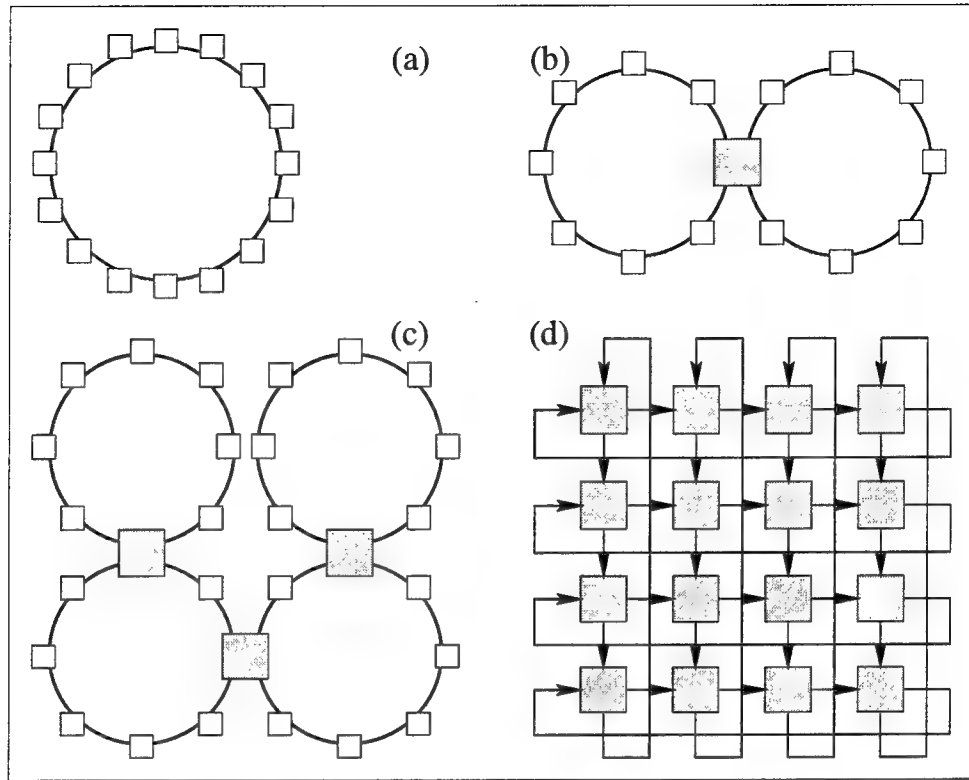


Figure 9: Simulated topologies: (a) Single ring with 16 nodes, (b) Two interconnected rings with 8 nodes/ring, (c) Four interconnected rings with 4 nodes/ring, (d) 2-D torus of dimension 4. Shaded box denotes a switch, white box denotes a simple node.

As can be seen from all the graphs that latency increases with increase in load and the flow control has a negative effect on the latency. In other words, flow control tends to increase the latency for the same system load. The reason being the flow control in SCI, in order to guarantee forward progress, may stop transmission from some nodes and thereby increases the average delay. Similar observation was made in [6] for a single ring. Our experimentation validates that the effect extends to interconnected rings as well. Another interesting observation that can be made from the graphs is the effect of switches over the topologies. A topology where all nodes are switches performs much better than the one where all nodes are simple nodes. Other topologies show intermediate performance in between these two extremes. This can be attributed to the ability of a switch to partially “isolate” a ring from the outer traffic. Moreover, introduction of switches results in smaller rings for the same number of nodes and this in turn results in smaller ring transfer delay<sup>2</sup>. This

<sup>2</sup>It is the time between the transmission of the first bit of a packet from the source to the ring and the reception of the last bit of the packet by the target from the ring.

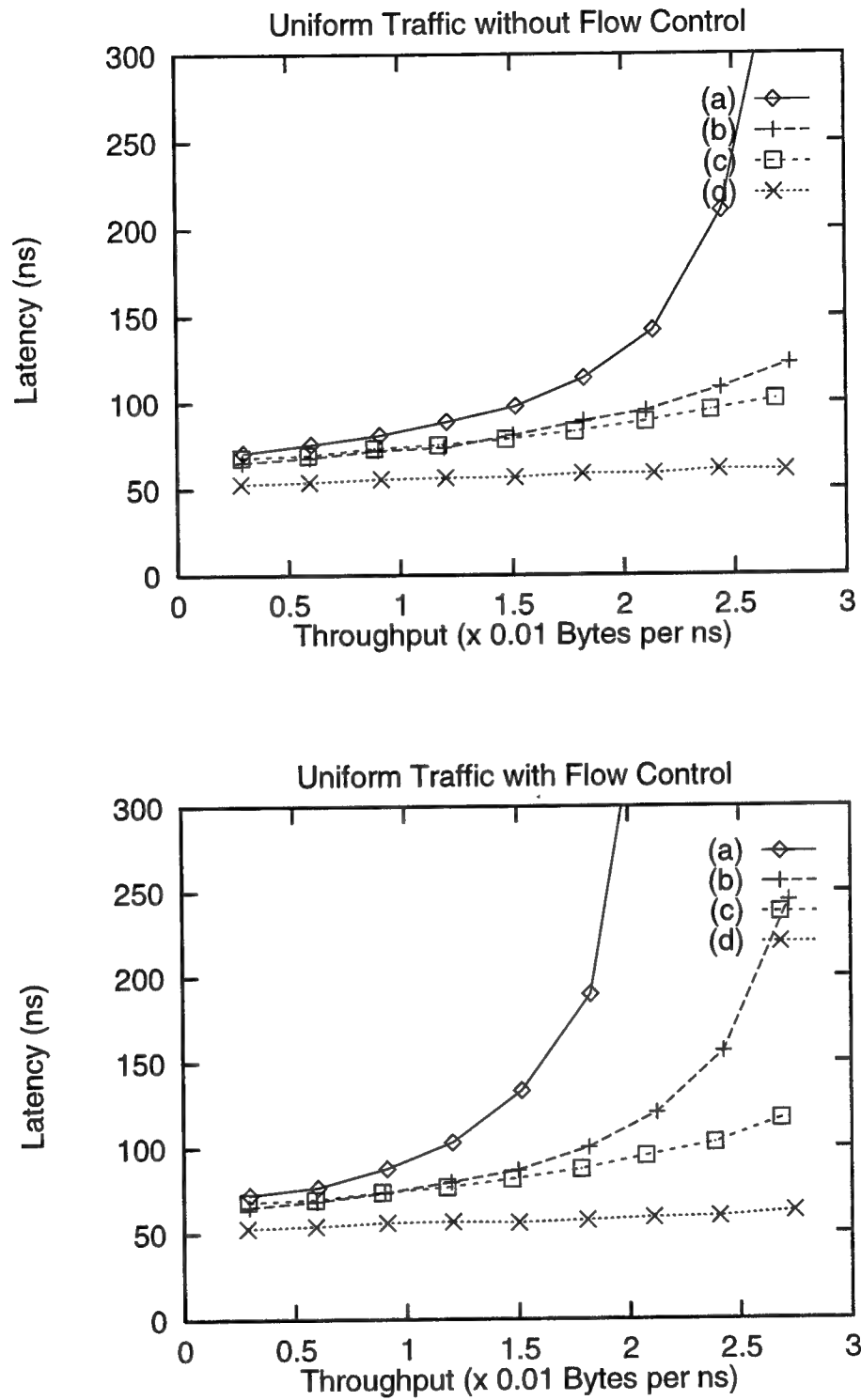


Figure 10: Uniform traffic with and without flow control. The index refers to the interconnected topology with reference to figure 9.

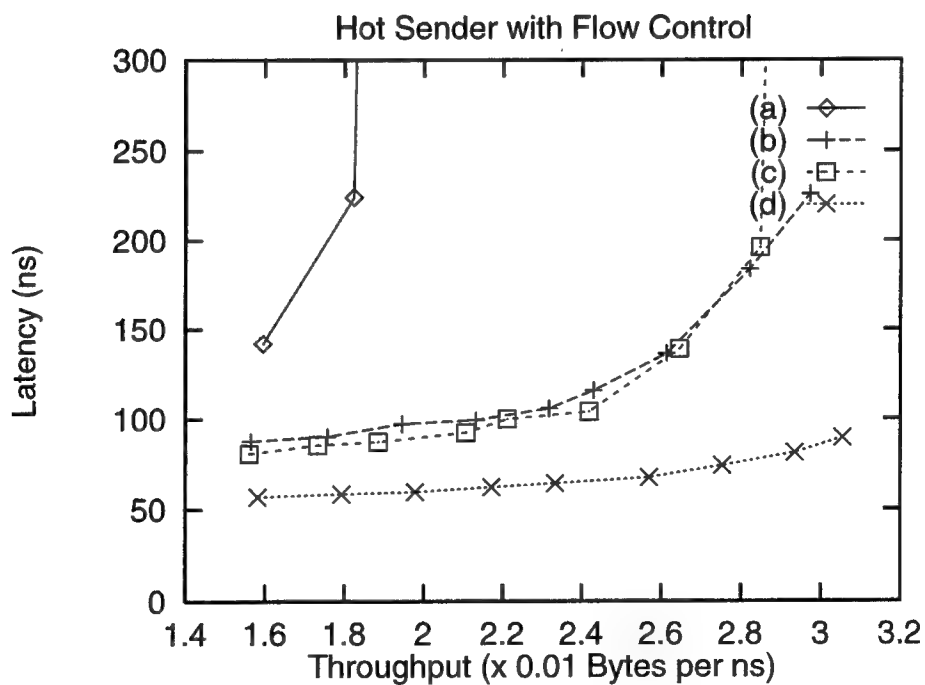
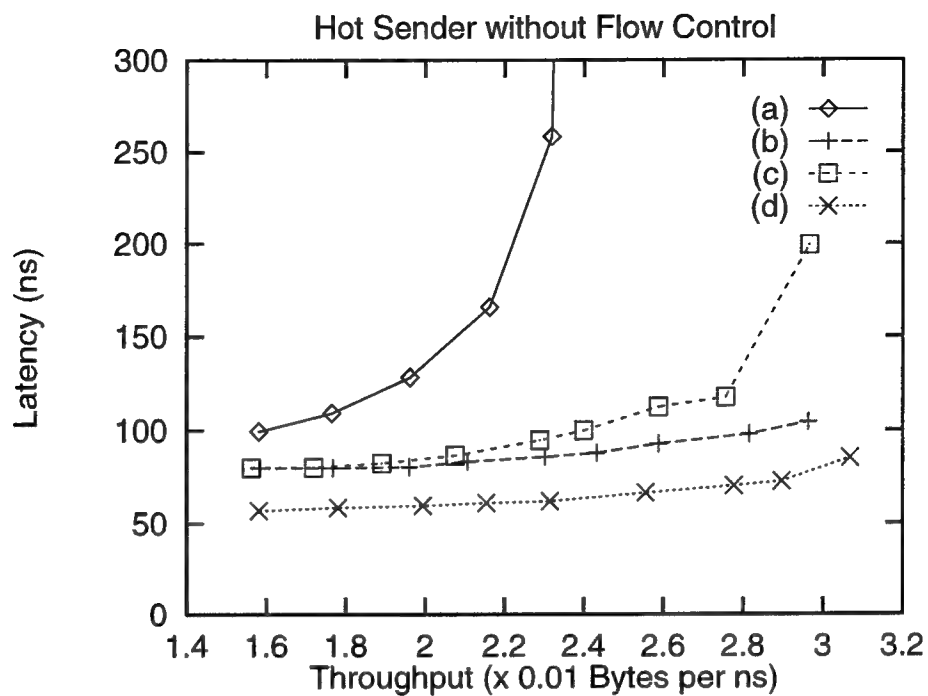


Figure 11: Hot sender with and without flow control. The index refers to the interconnected topology with reference to figure 9.

twofold effect reduces the latency drastically. The following paragraph summarizes our results.

**Summary of Results:** We conclude the following from the graphs:

1. Network topology plays a critical factor on system performance. From best to worst: (1) 2-D torus, (2) four interconnected rings, (3) two interconnected rings, and (4) single ring.
2. Flow control has a negative effect on the average system performance. The impact is dependent of topology. From least negative impact to most negative impact is: (1) 2-D torus, (2) four interconnected rings, (3) two interconnected rings, and (4) single ring.
3. Performance impact of traffic locality (uniform vs hot sender) is dependent on topology. From least dependent to most is: (1) 2-D torus, (2) four interconnected rings, (3) two interconnected rings, and (4) single ring.

## 5 Conclusion and Future Research

In this document we reported the work conducted on Scalable Coherent Interface during the summer employment of the author at the Wright Laboratory. The main thrust of the work was to study the performance of SCI in an interconnected topology. In order to carry it out, first a detailed simulation model of SCI node as well as switch were developed. The design and implementation were kept flexible enough to simulate different internal node architectures and interconnected topologies. We detailed the simulation model adopted and results obtained from the simulations. For ease of readers understanding we also presented a brief description of SCI.

The SCI protocol offers several attractive features. For example, it can provide high-speed point-to-point interface, can potentially allow larger distance between nodes. Also it is highly scalable, versatile and retrofitable. Because of the cost and performance potential offered by the SCI concept, it has gained popularity in the avionic applications in particular and many weapon system applications in general. Most of these applications demand real-time guarantee on message delivery. However, SCI as it stands today, can guarantee forward progress, but not latency. There is an on going effort to modify SCI to handle real-time traffic. Several features of SCI are identified that are not suitable for real-time application. The major problem identified by the author is due to the buffer-insertion feature of SCI. Because of this, SCI inherently introduces high variability in ring transfer time which makes the source-to-target delivery time non-deterministic. Several other hurdles identified (independently by the author and [3, 5]) are the FIFO queueing discipline, insufficient number of priority levels, etc. We are currently conducting research to introduce minimal modifications to the current SCI standard in order to make it suitable for real-time applications.

## References

- [1] B. W. Abeyesundara and A. E. Kamal. High-Speed Local Area Networks and Their Performance: A Survey. *ACM Computing Surveys*, 23(2):221-264, June 1991.
- [2] M. H. Davis and R. J. Kreutzfeld. An SCI-Based Architecture for Real-Time Avionics Processing. In *Proceedings of the IEEE*, pages 279-286, 1995.
- [3] D. B. Gustavson, B. E. Stewart, and D. L. Anderson. *SCI/RT: D0.13*, November 1992.
- [4] IEEE Computer Society. *IEEE Standard for Scalable Coherent Interface (SCI)*, August 1993. IEEE Standard 1596-1992.
- [5] D. James. *Draft Proposal for Real-Time Transactions on SCI*. Apple Computer Inc., April 1995. Version 0.27.
- [6] S. L. Scott, J. R. Goodman, and M. K. Vernon. Analysis of the SCI Ring. Technical Report 1055, Computer Science Department, University of Wisconsin-Madison, November 1991.

**Surface Resistance of High Temperature Superconductor  
Films Using Dielectric Resonator Measurements**

**Krishna Naishadham**

Associate Professor

Department of Electrical Engineering

Wright State University

Dayton, OH 45435

Final Report for:

Summer Faculty Research Program

Wright Laboratory

Sponsored by:

Air Force Office of Scientific Research

Bolling Air Force Base, DC

and

Materials Directorate

Wright Laboratory, OH

September 1995



# Surface Resistance of High Temperature Superconductor Films Using Dielectric Resonator Measurements

Krishna Naishadham

Associate Professor

Department of Electrical Engineering

Wright State University

Dayton, OH 45435

## Abstract

High temperature superconducting (HTS) materials are finding several applications in microwave circuits and devices. It has therefore become necessary and important that microwave characterization of HTS films be performed in WL/MLPO. This research explores the use of dielectric resonators to measure the surface resistance of HTS thin films at microwave frequencies. HTS dielectric resonators, besides offering a method of microwave characterization, are potentially useful as components in microwave systems (*e.g.*, as high-Q filters). We have measured the scattering (S) parameters of loop-coupled cylindrical dielectric resonators containing conductor films, as a function of temperature between 19 GHz and 40 GHz, using an automated network analyzer. By appealing to microwave circuit theory, an efficient algorithm has been developed to extract the surface resistance of the film from the measured S-parameters. The algorithm has been implemented on a 486-PC, and validated for copper films with independent measurements. Potential usefulness of the algorithm to characterize HTS films is discussed.

# Surface Resistance of High Temperature Superconductor Films Using Dielectric Resonator Measurements

Krishna Naishadham

## I Introduction

The discovery of high temperature superconductivity in LaBaCuO at 30K by Bednorz and Müller (1986) [1], and in YBaCuO (YBCO) at temperatures above 90K by Chu and several others (1987) [2], has significant impact on the design of microwave systems. Because of extremely small losses (or high Q-factor), low noise, low power consumption, potential for circuit miniaturization, high critical current densities, and uniform electrical behavior over a wide temperature range, high temperature superconductor (HTS) materials are becoming increasingly useful in aerospace industry, where size, weight and performance have high priority. Several designs of passive HTS microwave circuits, such as ultra low-loss transmission lines, high-Q microwave filters, high-gain antenna arrays, etc., have been reported [3]. In addition, HTS materials exhibit non-linear field effects at the macroscopic level (*e.g.*, Josephson tunneling effect), which made possible a number of active devices, such as field effect transistor (FET) and heterojunction bipolar transistor (HBT), operating with improved performance over their room-temperature normal conductor counterparts [4].

Most of the microwave applications of HTS materials employ thin or thick-film technology, in contrast to bulk materials. Microwave characterization of HTS thin films is essential to the development of these microwave applications. Also, the HTS microwave technology crosses several disciplines, because the successful development of an electronic system requires collaboration between material scientists and physicists, who are primarily involved in fabrication and characterization of the films, and microwave engineers, whose expertise involves the design and performance optimization of HTS microwave circuits. The most important thin-film HTS parameter for microwave applications is the surface resistance,  $R_s$ , which determines the dissipation in microwave devices, and hence the Q. This collaborative research with WL/MLPO addresses how the surface resistance can be measured accurately with a microwave network analyzer using dielectric resonators [5]. The author's objective of the research is to develop a robust and accurate method for extracting the surface resistance from narrow-band measurements of scattering ( $S$ ) parameters of the films. Therefore, the details of measurement procedure are not dealt with in this report. The interested reader is referred to [6] - [8] for these details. The reflection and transmission measurements on the films were carried out by Dr. Eric Moser at WL/MLPO at several frequency bands in the range 19 GHz to 40 GHz, and at several temperatures. The measurement procedure is still evolving, and technical difficulties involved in the accurate microwave characterization of small-area HTS

thin films, grown over sapphire substrates by pulsed laser deposition (PLD) at WL/MLPO, are not completely resolved. For this reason, the results presented in this report should be considered as preliminary, and are primarily concerned with establishing feasibility for accurate extraction of surface resistance from measurements on thin copper and niobium plates of moderate surface area.

The report is organized as follows. In the next section, a brief survey of two existing methods to extract surface resistance from dielectric resonator measurements of S-parameters is presented, and the limitations of the two methods are discussed. In the third section, we describe an equivalent circuit representation of the dielectric resonator cavity and coupling mechanism, and discuss how the reflection and transmission coefficients of the equivalent circuit circumscribe circles in the complex plane as a function of frequency. These circles, known as Q-circles [5], [9], characterize dielectric resonators in a narrow frequency range around the resonant frequency. Subsequently, in Sec. IV, we introduce a new method, called herein as the curve-fitting procedure (or CFP for short), to accurately fit Q-circles to the measured data. The CFP is validated with simulated examples (Sec. V) and measured examples (Sec. VI). Sec. VI also discusses how the limitations of the measured data on superconductors can be projected by the CFP as *unreliable data*. The important conclusions of this research, as well as some details on the follow-up effort to meet the HTS microwave characterization needs of WL/MLPO, are summarized in Sec. VII.

## II Dielectric Resonator Measurements

Cavities containing dielectric resonators are very useful in measuring the surface resistance of HTS thin films. Fig. 1 shows a sapphire dielectric resonator with HTS end caps [10]. The resonator can be either free-standing, or enclosed in a metallic package (cavity) as in Fig. 1. Energy is coupled into and out of the cavity through two coupling loops. By proper design and placement of these loops, one can ensure that only the dominant  $TE_{011}$  mode is excited within the resonator. Because the fields are well-trapped in the dielectric and within a small cylindrical region outside the sapphire, the losses in the resonator system can be minimized, with the result that extremely large Q's (of the order of a million) can be measured.

In order to measure the surface resistance of YBCO HTS films over a wide frequency range, five cylindrical sapphire pucks of different diameters were fabricated. The HTS films were placed non-destructively on either end of the sapphire puck, and their resonant frequencies as well as the insertion loss around the resonant frequency were measured by monitoring microwave transmission through the resonator. The loops were made using coaxial cables, and connected to the 50  $\Omega$  test

ports of HP8722 network analyzer for automated measurement.

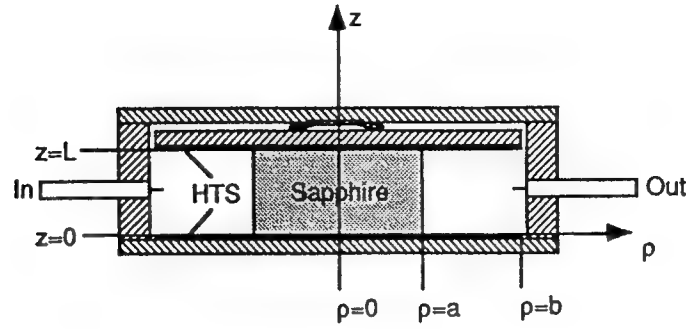


Fig. 1. Cross-sectional view of the HTS-sapphire-HTS dielectric resonator [10].

The fields trapped within the dielectric (Fig. 1) are oscillatory and described by Bessel functions of the first kind,  $J_n(x)$ . The evanescent fields along the radial direction are specified by (decaying) Bessel functions of the second kind,  $K_n(x)$ . These fields and their behavior are analyzed in [5]. By imposing boundary conditions on the tangential fields at the dielectric interface  $\rho = a$ , one obtains the transcendental equation

$$\frac{J_1(\xi_1 a) K_0(\xi_2 a)}{\xi_1 a} + \frac{K_1(\xi_2 a) J_0(\xi_1 a)}{\xi_2 a} = 0 \quad (1)$$

where  $\xi_1$  and  $\xi_2$  are radial wavenumbers in the dielectric ( $\rho \leq a$ ) and air ( $\rho \geq a$ ), respectively. In order to solve eq. (1) numerically for the resonant frequency, we provide initial guesses of the frequency and  $\xi_1 a$  and calculate

$$\xi_2 a = \pi \sqrt{(a/L)^2 - (2a/\lambda)^2} \quad (2)$$

where  $\lambda$  is the operating wavelength. The resonant frequencies for five different puck diameters were computed from (1) using Mathcad, and are listed in Table 1. The dielectric constant of sapphire is assumed as  $\epsilon_r = 9.3$ . These resonant frequencies are in good agreement with measured values.

**Table 1. Computed Resonant Frequencies of Sapphire Dielectric Resonators.**

Radius $a$	Length $L$	Res. Freq. (GHz)
0.09"	0.137"	37.276
0.095"	0.137"	35.888
0.1"	0.137"	34.628
0.12"	0.137"	30.607
0.15"	0.137"	26.587
0.25"	0.137"	20.364

We have also investigated the evanescent field decay along the radial direction external to the dielectric puck, to provide insight into determining a puck size that would ensure a small perturbation of the resonant mode of the dielectric resonator. The details are not presented in this report for brevity.

Two methods have been in use to extract the unloaded Q-factor of the dielectric resonator, namely, Ginzton method [11] and Kobayashi's method [12]. Both of these methods are applicable to the processing of S-parameters measured by the microwave network analyzer, and will be briefly discussed next. The limitations of these two methods will also be presented.

## **II-A Ginzton Method**

The network analyzer measures the reflection and transmission coefficients at the two ports connected to the coupling loops (see Fig. 1). The analyzer is equipped to locate frequency markers at arbitrary frequency points on these responses. For a resonant cavity, the magnitude of the insertion loss ( $S_{21}$  expressed in dB) follows the peaked behavior shown in Fig. 2 near the resonant frequency. Ginzton's method [11] entails the observation of response at only two or three frequency points for determination of the loaded Q-factor,  $Q_L$ , which can be calculated as

$$Q_L = \frac{f_L}{\Delta f}. \quad (3)$$

Here,  $f_L$  is the loaded resonant frequency,  $\Delta f$  is the frequency spread between 3 dB points.

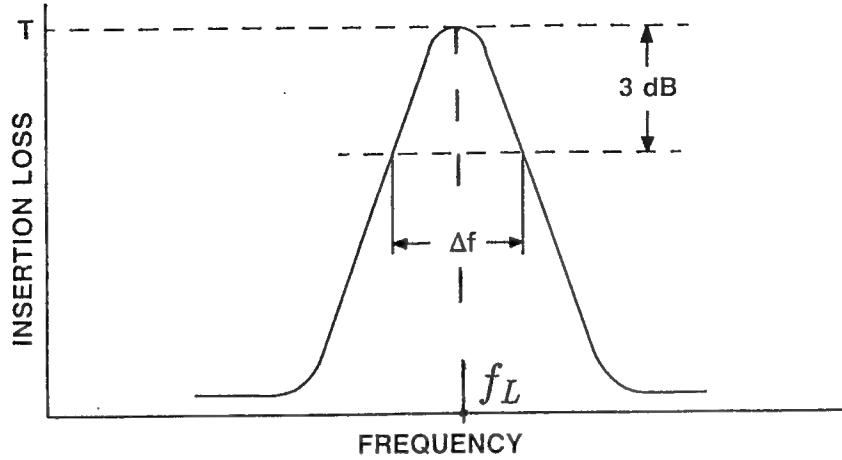


Fig. 2. Resonant curve measurement of the loaded Q-factor of a dielectric resonator.

Ginzton's method employs measured data from a very few frequencies around the loaded resonant peak, and thus, suffers from the following limitations. First, the unloaded Q-factor cannot be readily calculated because the magnitude response lacks information on the coupling coefficient, which determines the proportion of source power actually coupled into the resonator. Second, if the data is either unsymmetric around the peak or corrupted by measurement noise, an extraction procedure based only on magnitudes may yield very unreliable results. The phase of the measured S-parameters becomes important in these situations.

## II-B Kobayashi Method

Kobayashi method [12] also employs an HTS dielectric resonator operating in the  $TE_{011}$  mode to determine the surface resistance of HTS films. The extraction procedure in Kobayashi's method improves upon the Ginzton method by providing the coupling coefficient, from which the unloaded Q-factor may be determined. Essentially, the loaded Q-factor is still computed from the resonant peak and the two 3 dB points, as in Ginzton's method (see (3)). Kobayashi, however, assumes that the input and output coupling coefficients are equal, and determines the unloaded Q-factor from the insertion loss,  $T$ , at the resonant frequency  $f_L$  (see Fig. 2):

$$Q_u = \frac{Q_L}{1 - T}, \quad (4)$$

$$T = |S_{21}| = |S_{12}| = \frac{2\kappa_c}{2\kappa_c + 1} \quad (5)$$

where  $\kappa_c$  is the coupling coefficient at either port.

We have found that Kobayashi's method requires fairly moderate coupling for accurate prediction of the unloaded  $Q$ . It is difficult to ensure that the loops are always correctly positioned for equal coupling, especially with the small resonator fixtures that we employ at higher frequencies. Since Kobayashi's method is also based on magnitude measurements on the resonant curve, it suffers from the same limitations as Ginzton's method, and hence, yields unreliable results in some practical situations.

### III Equivalent Circuit Modeling

A resonator, in principle, has many modes with different resonant frequencies. However, if attention is focused on the dominant mode, which is the only one typically excited, the dielectric resonator can be conveniently represented by a parallel tuned circuit [9]. Thus, microwave circuit theory can be employed to formulate a robust extraction algorithm for the determination of the unloaded  $Q$ -factor. Unlike Ginzton's and Kobayashi's methods, such an algorithm would utilize both magnitude and phase of the measured  $S$ -parameters.

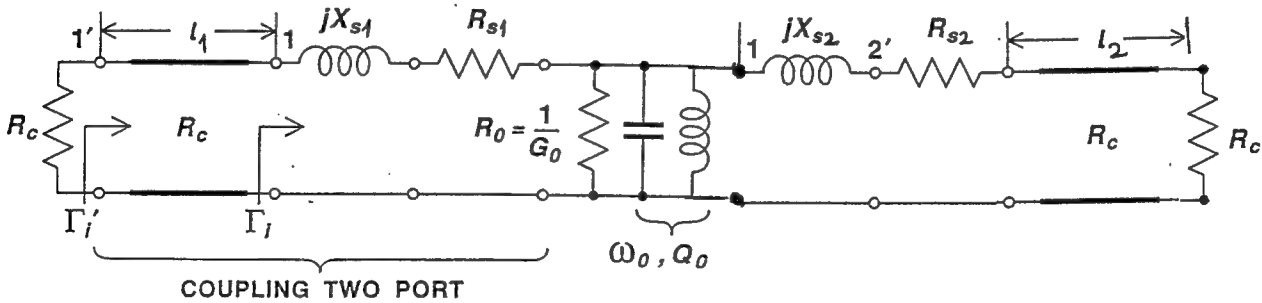


Fig. 3. Equivalent circuit of the dielectric resonator configuration, including the input and output coupling loops.

An equivalent circuit of the dielectric resonator configuration of Fig. 1, including the coupling loops, is shown in Fig. 3. The resonator is completely specified by the resonant frequency  $\omega_0$ , the unloaded  $Q$  factor  $Q_0$ , and the conductance  $G_0$  (or the resistance  $R_0$ ). The input and output

coupling loops are each modeled by a series resistance  $R_s$  and reactance  $X_s$ . The series resistance accounts for the power dissipated in the coupling loop. The series reactance includes the reactance of the loop, and also encompasses the influence of all higher-order resonant modes with distant resonant frequencies. This influence is usually negligible. Therefore, the equivalent circuit is valid only near the first (fundamental) resonance. The analyzer is connected to the loops by means of two transmission lines with characteristic impedance  $R_c$ . For modeling purposes, these lines are assumed to have lengths  $\ell_1$  and  $\ell_2$ , respectively, at the input and output ends. In practice, since the measurement reference planes cannot be located with any reasonable certainty, these lengths are undeterminable.

The unloaded admittance of the resonator is calculated as

$$Y_0 = \frac{1}{R_0} \left[ 1 + jQ_0 2 \frac{\omega - \omega_0}{\omega_0} \right] \quad (6)$$

where  $\omega$  is the operating frequency and  $\omega_0$  is the unloaded resonant frequency. This admittance does not consider the external loading of the coupling loops. The connecting transmission lines also affect the magnitude and phase of the resonator admittance. The loaded admittance of the resonator is given by

$$Y_L = \frac{1}{R_0} (1 + \kappa_1 + \kappa_2) \left[ 1 + jQ_0 2 \frac{\omega - \omega_L}{\omega_0} \right] \quad (7)$$

with the coupling coefficients due to external loading calculated as

$$\kappa_1 = \frac{(R_c + R_{s1})R_0}{(R_c + R_{s1})^2 + X_{s1}^2}, \quad \kappa_2 = \frac{(R_c + R_{s2})R_0}{(R_c + R_{s2})^2 + X_{s2}^2}. \quad (8)$$

Here,  $\omega_L$  is the loaded resonant frequency. It is observed from (8) that each coupling coefficient may be written as

$$\kappa_k = \kappa_k^l + \kappa_k^c, \quad k = 1, 2, \quad (9)$$

$$\kappa_k^l = \frac{R_c R_0}{(R_c + R_{sk})^2 + X_{sk}^2}, \quad \kappa_k^c = \frac{R_{sk} R_0}{(R_c + R_{sk})^2 + X_{sk}^2}. \quad (10)$$



Notice that superscript  $l$  denotes coupling associated with the transmission line, whereas superscript  $c$  denotes that caused by the loop. Physically, each coupling coefficient equals the ratio of power dissipated in the external component to power dissipated in the resonator. Using standard circuit theory [9], the input impedance at each port can be calculated as

$$Z_1 = Z_{e1} + \frac{1}{Y_o + Y_{e2}}, \quad Z_2 = Z_{e2} + \frac{1}{Y_o + Y_{e1}} \quad (11)$$

$$Z_{ek} = \frac{1}{Y_{ek}} = R_c + R_{sk} + X_{sk}, \quad k = 1, 2. \quad (12)$$

The port reflection coefficients are then given by

$$S_{kk} \equiv \Gamma_k = \Gamma_{dk} + \frac{2\kappa_k^l}{1 + \kappa_1 + \kappa_2} \frac{e^{j\gamma_k}}{1 + jQ_L 2 \frac{\omega - \omega_L}{\omega_0}} \quad (13)$$

with the unloaded and loaded Q factors related by

$$Q_0 = Q_L(1 + \kappa_1 + \kappa_2). \quad (14)$$

In the limit as the resonator is detuned to an extremum on either side of  $\omega_L$ , it is evident from (13) that the reflection coefficient approaches a value  $\Gamma_{dk}$  given by

$$\Gamma_{dk} = \frac{R_{sk} + jX_{sk} - R_c}{R_{sk} + jX_{sk} + R_c}. \quad (15)$$

The transmission coefficient also can be derived by appealing to circuit theory, and is given by

$$S_{21} = S_{12} = \frac{2\sqrt{\kappa_1^l \kappa_2^l}}{1 + \kappa_1 + \kappa_2} \frac{e^{-j\phi}}{1 + jQ_L 2 \frac{\omega - \omega_L}{\omega_0}}. \quad (16)$$

The phase angles  $\gamma_k$  and  $\phi$  are functions of loop parameters  $R_s$  and  $X_s$ , and are given by

$$\gamma_k = 2 \arctan \frac{X_{ks}}{R_c + R_{ks}}, \quad (17)$$

$$\phi = \arctan \frac{X_{1s}}{R_c + R_{1s}} + \arctan \frac{X_{2s}}{R_c + R_{2s}}. \quad (18)$$

### III-A Q Circles

As the frequency increases, the reflection and transmission coefficients describe circles in the complex plane, known as *Q circles* [9]. The unloaded Q-factor can be accurately computed from the center and diameter of the Q-circle. As an example, Fig. 4 shows the Q circle for  $S_{11}$ , plotted from (13), for the case:

$$\begin{aligned} \frac{R_o}{R_c} &= 2, \quad Q_0 = 1000, \\ \frac{R_{s1}}{R_c} &= 0.2, \quad \frac{R_{s2}}{R_c} = 0.4, \quad \frac{X_{s1}}{R_c} = 0.5, \quad \frac{X_{s2}}{R_c} = 1.5, \\ \beta_0 l_1 &= \beta_0 l_2 = 40 \text{ deg.}, \quad f_0 = 1 \text{ GHz}, \end{aligned}$$

where  $\beta_0$  is the free space wavenumber at  $f_0$ . The circle is obtained by plotting  $S_{11}$  over a frequency band of  $\pm f_0(3/Q_0)$  around the unloaded resonant frequency. A few observations can be made on the Q-circle by examining (13). The diameter of the circle is given by

$$d_{11} = \frac{2\kappa_1^l}{1 + \kappa_1 + \kappa_2}. \quad (19)$$

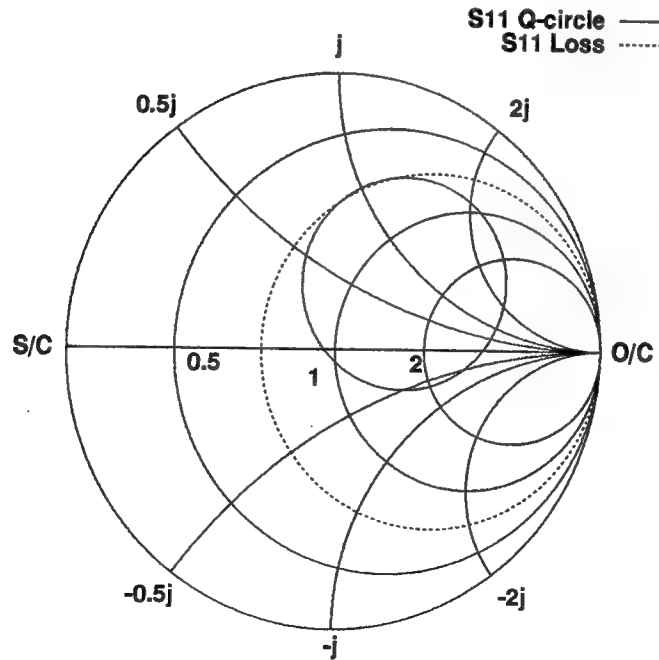


Fig. 4. Q-circle for the simulated reflection coefficient data.

The frequency at which the reflection coefficient is minimum is the loaded resonant frequency,  $f_L$ ,

and corresponds to a point on the circle which is closest to the center of the polar grid. The diametrically opposite extreme on the Q-circle denotes the detuned reflection coefficient. If the input coupling loop is lossless ( $R_{s1} = 0$ ), the detuned coefficient occurs on the rim of the Smith chart (see (15)). In the lossy case, the energy coupled into the resonator is reduced by the dissipation in the loop, with the result that the Q-circle is tangential to a circle, known as the *coupling loss circle*, at the detuned point. The loss circle is shown by the dashed curve in Fig. 4. Similar Q-circles and loss circles can be drawn for  $S_{22}$  and  $S_{12}$ . The dependence of diameters of all these circles on the coupling coefficients provides a system of equations to be solved for these coefficients. Details will be provided in the appendix.

The Q-circles are usually not smooth for measured data because of extraneous noise and other limitations of the measurement system. However, because of the physical reasoning that the measured S-parameters of loop-coupled resonators must follow the Q-circles as a function of frequency [9], a curve-fitting procedure may be used to fit a smooth circle through the data.

#### IV Least Squares Curve Fitting Procedure

The reflection and transmission coefficients given by eqs. (13) and (16) are of the form

$$w_i = \frac{a_1 t_i + a_2}{1 + a_3 t_i} \quad \text{with} \quad t_i = 2 \frac{f_i - f_L}{f_0}. \quad (20)$$

Eq. (20) may be viewed as a fractional linear transformation mapping the normalized frequency variable,  $t_i$ , to the space spanned by  $w_i$ . The complex transformation constants  $a_1$ ,  $a_2$  and  $a_3$  are to be determined from the set of  $i$  measurements,  $f_i, w_i$ ,  $i = 1, 2, \dots, N$ , where  $w_i$  denotes the reflection or transmission constant at the frequency  $f_i$ . The functional dependence of these constants on physical parameters of the resonator may be determined by comparing the right hand side of (13) or (16) with that of (20). For example, from (16) we obtain

$$a_1 = 0, \quad a_2 = \frac{2\sqrt{\kappa_1^l \kappa_2^l} e^{-j\phi}}{1 + \kappa_1 + \kappa_2}, \quad a_3 = jQ_L. \quad (21)$$

Since the measured data is overdetermined, the problem of determining the transformation constants may be cast in the weighted least squares format [9]

$$[C][a] = [q] \quad \text{with} \quad C_{ij} = \langle e_i | P | e_j \rangle, \quad i, j = 1, 2, 3. \quad (22)$$

Note that the three independent measurements of  $S_{12}$ ,  $S_{22}$  and  $S_{11}$  yield three distinct linear systems exemplified in (22), or equivalently, three coefficient vectors  $|a\rangle$ . Each vector  $|a\rangle$  contains the three constants  $a_i$ , and  $q_i = \langle e_i | P | f \rangle$  where the vectors  $|e_i\rangle$  and  $|f\rangle$  are completely specified by the set of  $N$  measurements:

$$|e_1\rangle = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix} \quad |e_2\rangle = \begin{bmatrix} 1 \\ 2 \\ \vdots \\ N \end{bmatrix} \quad |e_3\rangle = \begin{bmatrix} t_1 w_1 \\ t_2 w_2 \\ \vdots \\ t_N w_N \end{bmatrix} \quad (23)$$

$$|f\rangle = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix}. \quad (24)$$

$[P]$  is a diagonal matrix of weights (*i.e.*,  $[P] = \text{diag}[p_i]$ ,  $i = 1, 2, \dots, N$ ), with  $p_i$  inversely proportional to the variance of the  $i^{\text{th}}$  measurement.

The linear system in (22) is iteratively solved for the vector  $|a\rangle$  until the root mean square (rms) error calculated from the transformation equation (20) is minimized below an upper bound. Details may be found in [9]. The coordinates of the center of each Q-circle in the complex plane, the detuned reflection coefficients, as well as the diameters  $d_{11}$ ,  $d_{22}$  and  $d_{12}$  of the Q-circles, can be uniquely determined from the corresponding vector  $|a\rangle$ . The loaded Q factor follows from  $a_3$  in (21). The appendix summarizes how one can obtain the coupling coefficients and the unloaded Q from these parameters determined by curve-fitting.

## V Simulated Example of Curve Fitting

We have determined the unloaded Q of the same resonator as described in Section III-A, with the data for curve-fitting provided by the data simulated using eqs. (13) and (16). Curve-fitting of the simulated data provides an intuitive validation of the computer program which has been written to implement the least squares CFP described in the previous section. In order to make the validation representative of actual measured data, simulated data lying within a small band of only  $\pm(f_0/Q_0)$  is input to the CFP program.

Fig. 5 shows the simulated and curve-fitted Q-circles for the reflection and transmission coefficients. In each case, the incomplete Q-circle shown as a polygon corresponds to the narrow-band

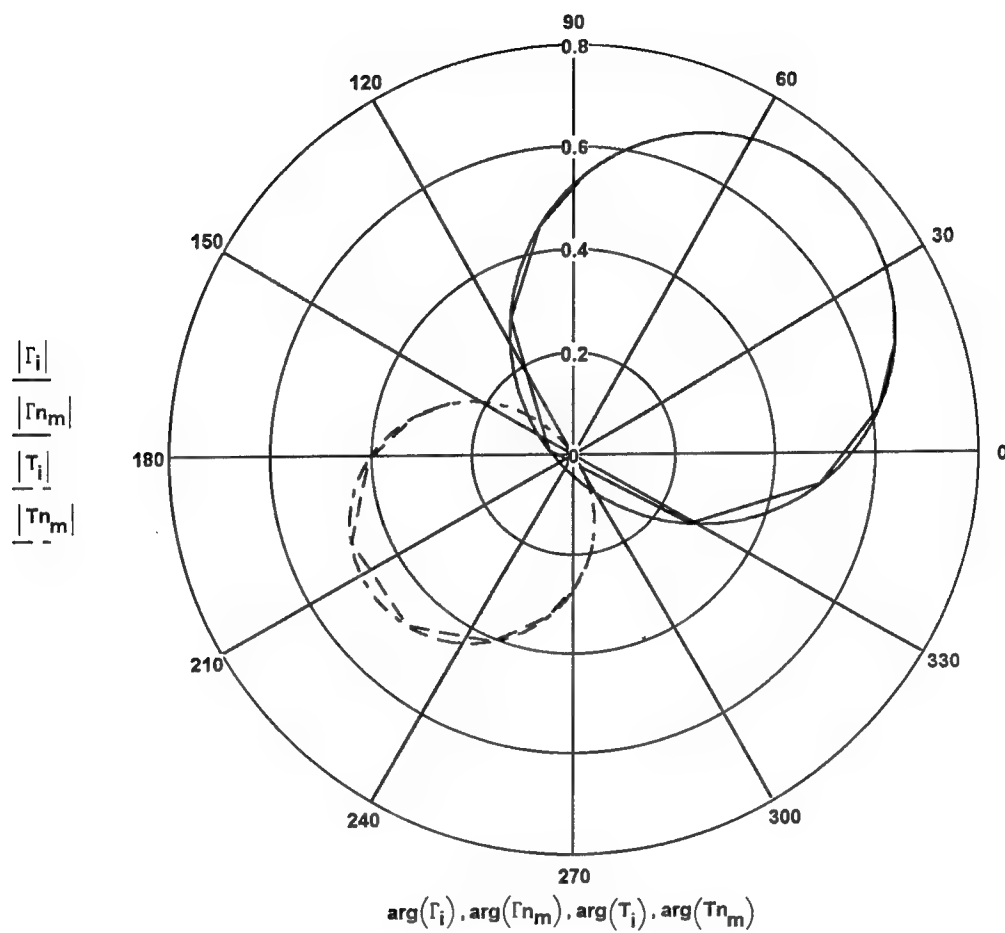


Fig. 5. Simulated and curve-fitted Q-circles for the example discussed in Sec. IIIA.

simulated data, while the complete circles are those obtained by CFP. Three iterations were used in the least squares CFP. The fitted circle coincides with the simulated data points. More importantly, using data from only 9 points, the CFP draws a smooth Q-circle through the simulated data. From the parameters of the fitted Q-circle, we have calculated  $\kappa_1 = 1.39012$ ,  $\kappa_2 = 0.65508$ ,  $Q_L = 325.128$ , which yield an unloaded Q of 990.1 (within 1% of the specified  $Q_0 = 1000$ ). We have performed several other validations with simulated data, and in each case, the parameters obtained from fitting were within  $\pm 1\%$  of the specified values. Thus, the accuracy of the program appears to be  $\pm 1\%$ .

## VI Measured Examples

A PC-based computer program has been written, using Mathcad, to implement the curve-fitting procedure for dielectric resonator measurements. The program accepts frequency-dependent measured reflection and transmission data and produces the unloaded Q and coupling coefficients. The losses within the coupling loops and connecting transmission lines need not be specified *a priori*. The program can determine these using the fitted Q-circles. Next, we present the salient results of CFP implementation on two sets of measured data: (a) a thin copper plate at room temperature, and (b) 1700 angstrom thin-film of niobium at 7K.

### VI-A Copper Plate

Fig. 6 displays in polar form the measured data and the fitted Q-circles for the input and output reflection coefficients, and the transmission coefficient. The measurements were taken in a narrow frequency band around 27 GHz. In order to see the accuracy of the fit, the input reflection coefficient and the transmission coefficient are replotted in Fig. 7 using a rectangular grid. The output port reflection coefficient is omitted for brevity. In the latter figure, the dashed line corresponds to the fitted curves, whereas the points denote the measurements. Five iterations of the least squares CFP are employed. It is evident that the measured data is dispersed around the fitted Q-circles. The data dispersion factor in the final iteration is found to be 0.04, indicating a reliable fit. From numerical experimentation, a dispersion factor less than 0.1 is found to produce reliable results as validated with published data. The unloaded Q determined by the fit is 4030, which is in good agreement with Kobayashi's measurement [12].

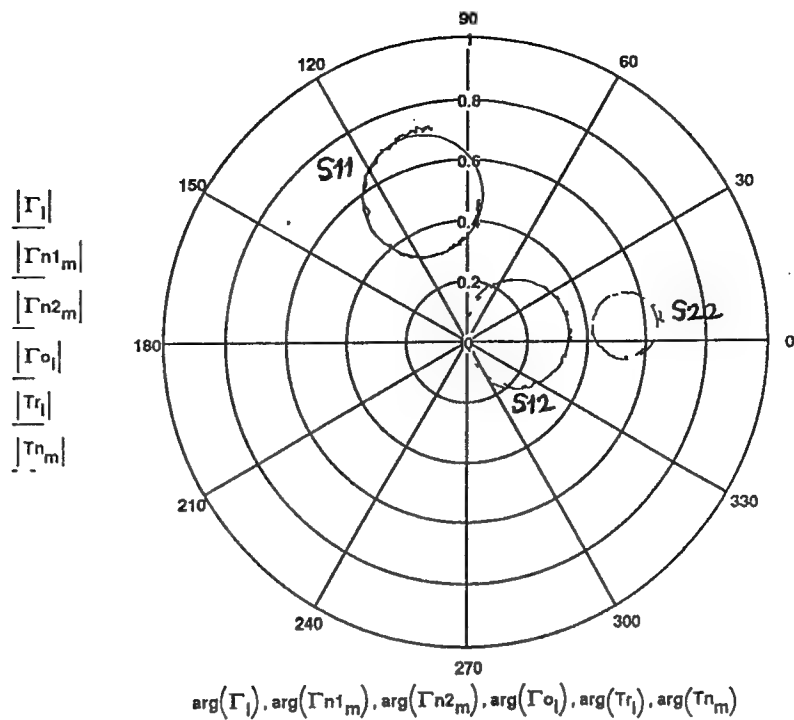


Fig. 6. Measured and curve-fitted Q-circles for the copper plate.

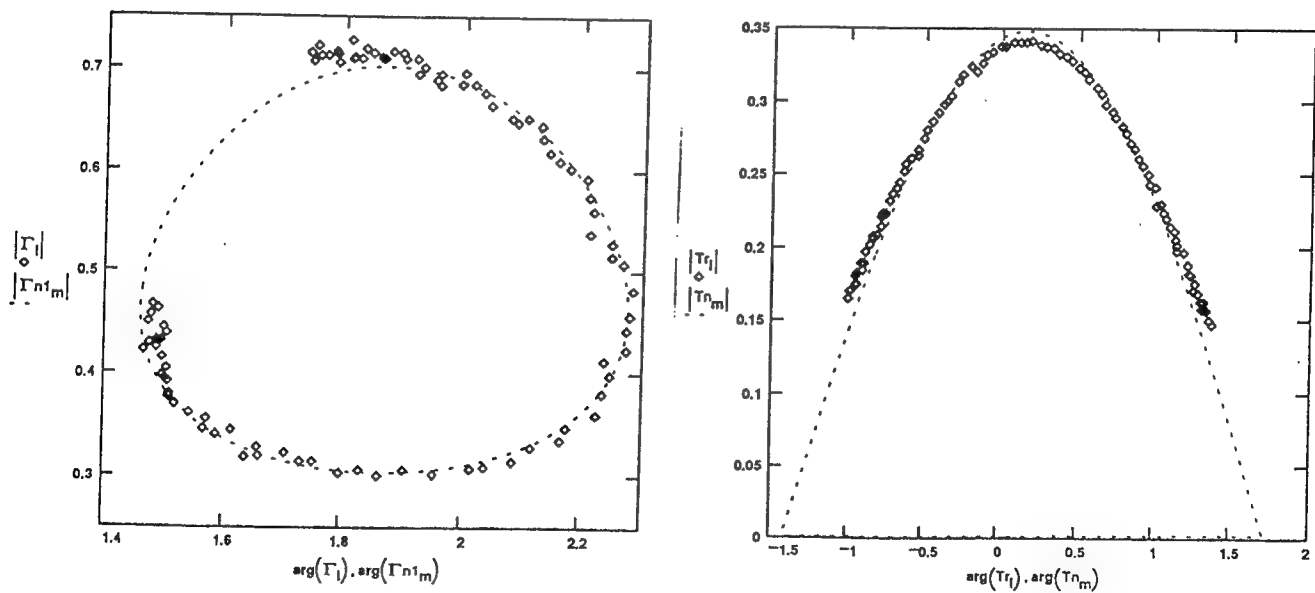


Fig. 7. Complex reflection and transmission coefficients for the copper plate replotted on a rectangular scale.

## VI-B Niobium Plate

Next we evaluate the measured data for a niobium plate of area  $10 \text{ mm}^2$  and a thickness of 1700 angstrom. Again, utilizing 5 iterations of the least squares CFP implementation, we arrive at the results displayed in Fig. 8. It is obvious that the fit is poor, as evidenced by the significant deviation between the fitted and the measured data. The unloaded  $Q$  calculated from the CFP program is 21,000 which is less than published measured data by a factor of 10! This is perhaps because of the fact that the niobium film was not polished, and because of measurement errors introduced by an inadequate fixture design. However, the CFP program correctly points out that the measured data is unreliable by consistently returning a value of data dispersion factor between 0.16 and 0.21. Therefore, the inaccuracy in the parameter extraction for this data set is caused by limitations of sample preparation and measurement fixture, and not the CFP program.

## VII Conclusions

An efficient algorithm, based on least squares minimization of the rms error between an assumed fractional linear transformation and the measured data, has been developed for the extraction of unloaded  $Q$  from dielectric resonator measurements. The algorithm has been implemented on a personal computer using Matchcad. This extraction procedure is more reliable and accurate than previous methods based on three-point resonant curve measurements (*e.g.*, Ginzton's and Kobayashi's methods), because a wide data sweep around the resonant frequency is used to fit the measured data. The extraction program has been applied to compute the unloaded  $Q$  of dielectric resonators consisting of copper and niobium plates, and has been validated with published data (copper). The capability of the program to detect flaws in the measurement or sample preparation has been demonstrated.

We conclude the report with a brief discussion on future work. Temperature-dependent measurements of HTS films need to be implemented in the program. Although we anticipate no insurmountable difficulties in extracting the data from HTS film measurements, mathematical formulation of the the curve-fitting linear transformation may need modification because of the extremely high  $Q$ 's of HTS resonators. Second, detailed analysis of the electromagnetic fields in the resonator and its fixture needs to be carried out in order to design a reliable fixture for *open* HTS dielectric resonators. Third, the CFP program needs to be implemented using a high-level language such as Fortran or C, so that large data sets encompassing measurements over several frequencies and temperatures, can be automatically processed by a single computer program. Mathcad lacks desirable programming features such as character I/O and iterative loops. Lastly, a graphical user



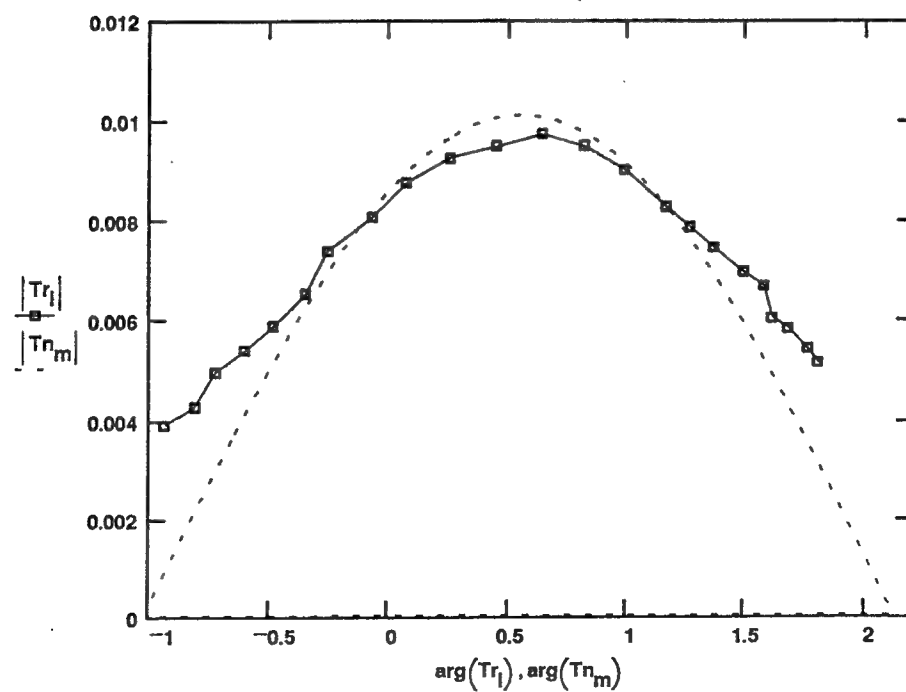
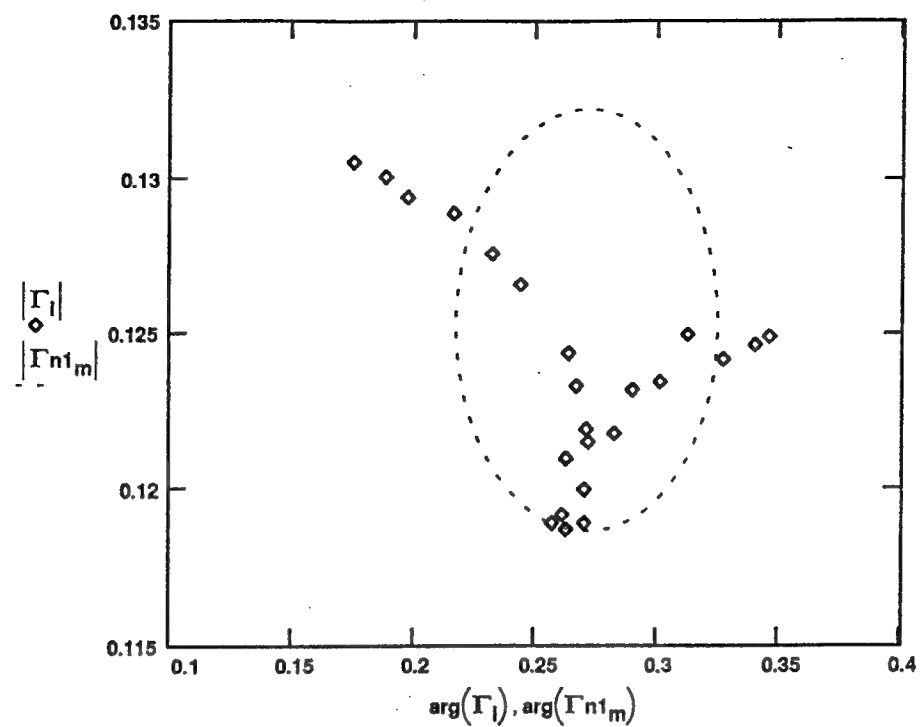


Fig. 8. Complex reflection and transmission coefficients for the niobium plate plotted on a rectangular scale.

interface needs to be developed for the high-level CFP program. It is anticipated that funds for this work will be provided by AFOSR SREP program and WL/MLPO.

## Appendix A

### Determination of Coupling Coefficients

With reference to Fig. 4, let  $d_{11}$  and  $d_{22}$  denote the diameter of the Q-circle for input and output reflection coefficients, respectively. The diameter of the corresponding coupling loss circle is denoted by  $d_{1c}$  and  $d_{2c}$ , respectively. The diameter of the transmission Q-circle is  $d_{12}$ . The diameter of each Q-circle follows readily from the corresponding transformation vector  $|a\rangle$  generated by the fitted curves [9]. The diameter of the loss circle is computed as

$$d_{kc} = \frac{d_{kk}[1 - |\Gamma_{dk}|^2]}{d_{kk} - (d_{kk}/2)^2 - |\Gamma_{dk}|^2 + |\Gamma_{ck}|^2}, \quad k = 1, 2 \quad (25)$$

where  $\Gamma_{ck}$  denotes the center of the corresponding reflection Q-circle. The various coupling coefficients are calculated as (see eqs. (9) and (10))

$$\kappa_1^l = \frac{d_{11}/2}{1 - d_{11} \left[ d_{1c}^{-1} - (d_{12}/d_{11})^2 d_{2c}^{-1} \right]} \quad (26)$$

$$\kappa_2^l = \kappa_1^l \left( \frac{d_{12}}{d_{11}} \right)^2 \quad (27)$$

$$\kappa_k^c = \kappa_k^l \left( \frac{2}{d_{kc}} - 1 \right), \quad k = 1, 2. \quad (28)$$

The unloaded Q factor follows from these coupling coefficients, the loaded Q (obtained using the curve-fitted Q-circle), and (14).

## References

- [1] J. G. Bednorz and K. A. Müller, "Possible high  $T_c$  superconductivity in the Ba-La-Cu-O system," *Z. fur Phys.*, vol. 64, p. 189, 1986.
- [2] M. K. Wu, J. R. Ashburn, C. W. Chu et al., "Superconductivity at 93K in a new mixed-phase Y-Ba-Cu-O compound system at ambient pressure," *Phys. Rev. Lett.*, vol. 58, p. 908, 1987.
- [3] *IEEE Trans. Microwave Theory Tech.*, Special Issue on Microwave Applications of Superconductivity, vol. 39, no. 9, pp. 1445-1594, Sep. 1991.
- [4] T. M. Klapwijk, D. R. Heslinga, and W. M. van Huffelen, "Superconducting field effect devices," in *Superconducting Electronics*, H. Weinstock and M. Nisenoff (eds.), Berlin Heidelberg: Springer-Verlag, pp. 385-408, 1989.
- [5] D. Kajfez and P. Guillon (eds.) *Dielectric Resonators*, Norwood, MA: Artech House, 1986.
- [6] Hewlett Packard Product Note No. 8510-3, "The measurement of both permittivity and permeability of solid materials," no. 5954-1535, Aug. 1985.
- [7] W. B. Weir, "Automatic measurement of complex dielectric constant and permeability at microwave frequencies," *Proc. IEEE*, vol. 62, no. 1, pp. 33-36, Jan. 1974.
- [8] L. P. Ligthart, "A fast computational technique for accurate permittivity determination using transmission line methods," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-31, no. 3, pp. 249-254, March 1983.
- [9] D. Kajfez, *Q Factor*, Oxford, MS: Vector Fields, 1994.
- [10] Z.-Y. Shen, *High-Temperature Superconducting Microwave Circuits*, Norwood, MA: Artech House, 1994.
- [11] E. L. Ginzton, *Microwave Measurements*, New York, NY: McGraw-Hill, 1957.
- [12] Y. Kobayashi, T. Imai, and H. Kayano, "Microwave measurement of temperature and current dependences of surface impedances for high- $T_c$  superconductors," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-39, no. 9, pp. 1530-1538, Sep. 1991.

**Robert P. Penno, Ph.D.**  
**Assistant Professor**  
**Department of Electrical Engineering**

**University of Dayton**  
**300 College Park**  
**Dayton, Ohio 45469-0226**

**Final Report for:**  
**Summer Faculty Research Program**  
**Wright Laboratory**

**Sponsored by:**  
**Air Force Office of Scientific Research**  
**Bolling Air Force Base, DC**  
**and**  
**Wright Laboratory**

**September, 1995**

**ANTENNA EFFECTS AND WL/AAAI'S  
ANTENNA WAVEFRONT SIMULATOR**

**Robert P. Penno, Ph.D.  
Assistant Professor  
Department of Electrical Engineering  
University of Dayton**

**Abstract**

A procedure for augmenting the accuracy of the Antenna Wavefront Simulator (AWFS) by improving the model of the antenna array simulated by the AWFS was studied. Use of a wire model of the antenna array accurately characterized the field pattern of the array, but did so with terminal currents not necessarily similar to those of the array under test. Other modeling techniques, such as the Finite Element Method (FEM), may be better suited to characterizing the type of antennas, typically microstrip patch antennas, to be evaluated for the AWFS. The wire model displayed some significant effects due to element mutual impedance. This was especially true of the reference element. The concept of perturbing the AWFS from the omnidirectional case seems valid. Also, use of the Geometrical Theory of Diffraction (GTD) to describe incidence from below horizon should be further investigated.

# ANTENNA EFFECTS AND WL/AAAI'S ANTENNA WAVEFRONT SIMULATOR

Robert P. Penno, Ph.D.

## Introduction

The Antenna Wavefront Simulator (AWFS) being developed by WL/AAAI-1 is an innovative concept rapidly becoming a reality. The goal of the AWFS is to provide a means of testing antenna array electronics packages in a dynamic, controlled laboratory environment<sup>1,2</sup>. Currently, such electronics packages are either simulated by computer model or evaluated via an anechoic chamber or open range field tests. The AWFS provides a real-time test platform that is inexpensive and compact. Moreover, it serves as an ideal supplement to the current design process, by affording the opportunity to test and modify in real time, before final testing in an open range environment.

Currently, the AWFS simulates the output of an incident signal upon an omnidirectional antenna element. Such an antenna element receives the same signal (magnitude and phase) independent of the direction of incidence. It is the array of such elements that provides the direction finding capability. To more accurately model the signal incident upon the electronics package, it is essential to incorporate into the AWFS a more accurate model of the antenna array elements being employed by the package. The AWFS would, then, provide signal outputs similar to those from an antenna array comprised of omnidirectional elements, but perturbed in magnitude and phase to incorporate the effects of the actual antenna elements. The purpose of this work was to analyze the Controlled Radiation Pattern Array (CRPA2) used with the AE-1 adaptive electronics package, which would lead to modifications of the existing AWFS to account for the use of a actual antenna instead of an ideal array of omnidirectional elements.

### Preliminary Model

Since the modeling of the CRPA2 is non-trivial, an intermediate, more primitive magnitude-only model was suggested. Such a model employed gain data obtained from actual range measurements, although no phase information was available or to be employed in the model. To test the viability of this model, a numerical model was created consisting of three coplanar, half wave dipoles, whose axes were rotated 60° from each other, similar to the orientation of the CRPA2. These dipoles were constrained to lie in the x-y plane. By using a wire grid method of moments code provided by Richmond<sup>3</sup>, the terminal currents of these dipoles for varying angles of incidence were determined. A comparison was made between the sum of the magnitudes of these terminal currents and the magnitude of the (complex) sum of the currents. As is expected from Schwartz' Inequality,

$$|Ae^{j\alpha} + Be^{j\beta} + Ce^{j\gamma}| \leq |A+B+C|,$$

where  $Ae^{j\alpha}$ ,  $Be^{j\beta}$ , and  $Ce^{j\gamma}$  are the terminal currents of the respective dipoles. Figures 1 and 2 display this variation as a function of azimuth angle for elevation angles of 0° and 90°, respectively. Using the magnitudes of these terminal currents (A, B, and C) to model the dipoles produces field patterns markedly distinct from the actual patterns derived from the magnitude of the sum of the complex current. In fact, it is reasonable to suggest that use of magnitude data only could actually introduce error in the modeling process. For this reason, implementation of a partial model was not pursued. Rather, a model that included both magnitude and phase information was required to accurately improve upon the existing omnidirectional element model.

### Modeling the CRPA2

As an improvement over the magnitude-only model, a wire antenna model was constructed to approximate the CRPA2. The CRPA2 is a seven element, cavity-backed, conformal, microstrip patch antenna array, whose central element is a quadruple-fed, circularly polarized patch antenna with an elaborate microstrip feed network. This element, the reference element, is used in other applications as a stand alone antenna.

Surrounding this reference element are six auxiliary elements, which are linearly polarized patch antennas whose principal axes are rotated 60° relative to adjacent auxiliary elements. No terminal current phase information is currently available for the CRPA2, while a limited amount of magnitude data is available. In the absence of complete terminal current information for the CRPA2, a wire model was developed whose field patterns approximated that of the CRPA2.

The wire model of the CRPA2 included ten wire structures (six dipoles and four folded monopoles), with their images, to approximate a ground plane. These dipole structures were input to a wire code that employs a method of moments solution to the Reaction Integral Equation (RIE) described by Richmond<sup>4</sup>. In this code all structures are represented by multi-segmented wires or wire-grids. Figure 3 describes the folded monopoles, and their images, used to depict the reference antenna. It was found that the height of the monopole greatly affected its radiation pattern. The feed network that sums the outputs of the four elements of the reference element to provide the circular polarization was not simulated by this code. Instead, the impedances of the individual elements were calculated and each terminal was loaded with the complex conjugate of its load impedance. To model the auxiliary elements, half wave dipoles (and their images) were used. The linear polarization of these auxiliary elements was obtained by an offset feed location. To simulate this with a dipole, the axis of the dipole was perpendicular to the polarization axis of the patch antenna. Figure 4 shows a layout of all the elements from a top view. The images are not shown, and are used only to simulate the effects of a ground plane. In all cases, the frequency employed in designing the wire model was 1.57542 GHz.

#### Determination of Matching Loads

The wire code treats the RIE as a system of linear, simultaneous equations of the form:

$$V_m = \sum Z_{mn} I_n$$

or, in matrix form,



$$[V] = [Z][I].$$

Here,  $V_m$  is the voltage across the  $m^{\text{th}}$  junction of two adjacent wire segments, while  $I_n$  is the current flowing through the  $n^{\text{th}}$  junction of two adjacent segments.

Thus,

$$[I] = [Z]^{-1}[V] = [Y][V],$$

where  $Y_{mn}$  is an element of the array,  $[Y]$ , and is equivalent to the short circuit admittance parameters. By placing a source of  $1\angle 0^\circ$  Volts at the terminals of one of the antenna elements with open circuits at the terminals of all the other elements, the current computed at this junction is identical to the short circuit admittance of the antenna element (as part of the antenna array). Repeating this calculation at each of the antenna element terminals fills the matrix,  $[Y]$ . Then, the impedance matrix,  $[Z]$ , is found by inverting this matrix as shown above. The elements of the matrix,  $[Z]$ , are the impedances that would be presented to a feed network. They are independent of the angle of incidence of the illuminating source; these impedances are a function of the geometry, and relative proximity of the elements in the antenna array. Given the impedance,  $Z_{22}$ , maximum power flow is obtained if the feed network to that terminal presents an impedance of  $Z_{22}^*$ , the conjugate of  $Z_{22}$ . The impedances of the four terminals of the reference antenna, as determined by the wire model, displayed an impedance of  $Z = 234 + j893 \Omega$ , while the impedances of the auxiliary elements were nominally  $29 + j66 \Omega$ . The currents at the four terminals are summed to present a right-handed, circularly polarized signal to the array. Changing this load to other than the conjugate match produced pattern results less similar to those of the CRPA2. Moreover, with pattern shape the highest consideration, the conjugate match provided the optimum loading.

#### Implementation into AWFS

The product of this computer study is an analysis of a wire model of the CRPA2 antenna. For a specified angle of source incidence, the terminal signals,  $I_n$  (above), presented to an adaptive array electronics package would be produced. These signals are

perturbations to the signals produced by an omnidirectional antenna array stimulated by the same source. The AWFS would employ these signals directly in a tabular form, i.e. the tables currently employed by the AWFS would be adjusted for a given angle of incidence to reflect the presence of the CRPA2, not an omnidirectional element.

#### Observations and Future Directions

The use of a wire model to approximate the CRPA2 patch antenna array was based upon approximating the field patterns, not terminal signals. Approximating the field patterns of the CRPA2 produced terminal signal strengths and impedances quite dissimilar to those typically found in the actual CRPA2. While the terminal impedances of the CRPA2 are not available, it is not likely that impedances such as those demonstrated by the reference element would be obtained on a  $50\Omega$  system. Moreover, the shape of the field patterns seemed to present a 50 dB difference in strength while retaining the same basic shape as is shown in Figures 5 and 6<sup>5</sup>. As such, these results proved to be less useful than expected.

Of more significance was an analysis of the mutual impedances of the wire model elements. The reference element, which consists of four elements as described above, displayed the greatest sensitivity to mutual impedance. In the presence of the auxiliary elements, the impedance at the terminal of the reference element was  $517.6\angle -8.95^\circ \Omega$ . Removing all of the auxiliary elements changed this impedance to  $836.12\angle 49.3^\circ \Omega$ . The auxiliary elements seemed to have very little impact upon the other auxiliary elements in the absence of the reference element. The impedance of one auxiliary element with or without the other auxiliary elements (without the reference element) was essentially unchanged. It is reasonable to suggest that the impact of mutual impedance in the CRPA2 would be similar, that is the reference element would be significantly impacted by itself and the auxiliary elements.

The concept of numerically modeling the antenna, and employing this model to perturb the signals from the omnidirectional case, seems to be credible. What is needed is a better model to approximate the antenna arrays likely to be presented to the AWFS for

evaluation. A likely candidate for such a model would be the finite element method. This approach has been used by various researchers<sup>6</sup> for cavity-backed, microstrip patch antennas, and could provide accurate terminal characteristics for illuminations above the plane of the array (horizon). Such a methodology might prove to be broad enough to handle most of the antenna arrays that would be presented to the AWFS.

It was observed in some data that for angles of incidence below horizon, the field patterns of the CRPA2 were very rough, displaying an almost interference-like return (See Figure 6). Though no work has been performed to corroborate this, it is speculated that this below horizon return is the result of electromagnetic waves hitting the bottom of the mounting of the CRPA2 (typically a barrel-shaped fuselage), inducing creeping waves which travel around the fuselage to the edge of the CRPA2. This is a well studied phenomena in high frequency scattering theory, for which a technique, the Geometrical Theory of Diffraction (GTD)<sup>7</sup>, has been developed. A logical extension to the work done here would be to incorporate GTD effects for angles of incidence below the horizon of the CRPA2 platform into the AWFS. GTD could be used in conjunction with other modeling techniques (e.g. FEM) to describe incidence from below the horizon.

---

<sup>1</sup> An Assessment of the WL/AAAI-4 Antenna Wavefront Simulator, Robert P. Penno, AFOSR # 94-0122, September, 1994

<sup>2</sup> Antenna Wavefront Simulator, Dana Howell, Bruce Rahn, WL/AAAI-1, WPAFB, Joint Services Data Exchange, October, 1994

<sup>3</sup> Computer Program for Thin-Wire Structures in a Homogeneous Conducting Medium, J.H. Richmond, The Ohio State University, Columbus, Ohio

<sup>4</sup> Radiation and Scattering by Thin-Wire Structures in the Complex Frequency Domain, J.H. Richmond, The Ohio State University, Columbus, Ohio

<sup>5</sup> Ron Whitsel, NRAD, Warminster, Pa.

<sup>6</sup> On the Analysis of a Stacked Patch Antenna, Stephen W. Schneider, WL/AAWW, WPAFB, Ohio

<sup>7</sup> "Antenna Theory and Design", W.L. Stutzman, G. A. Thiele, Wiley & Sons, 1981, Chapter 9

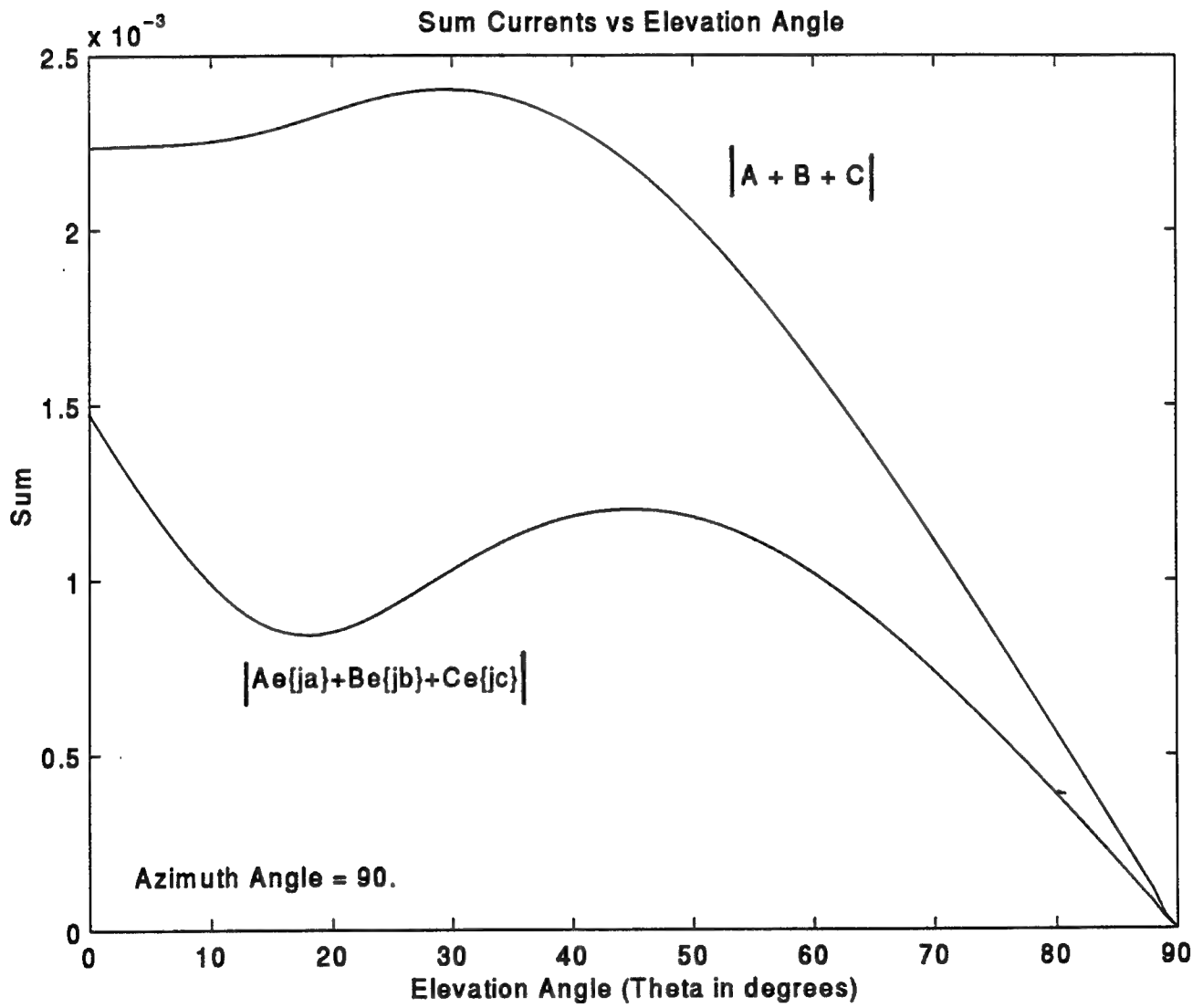


Figure 1

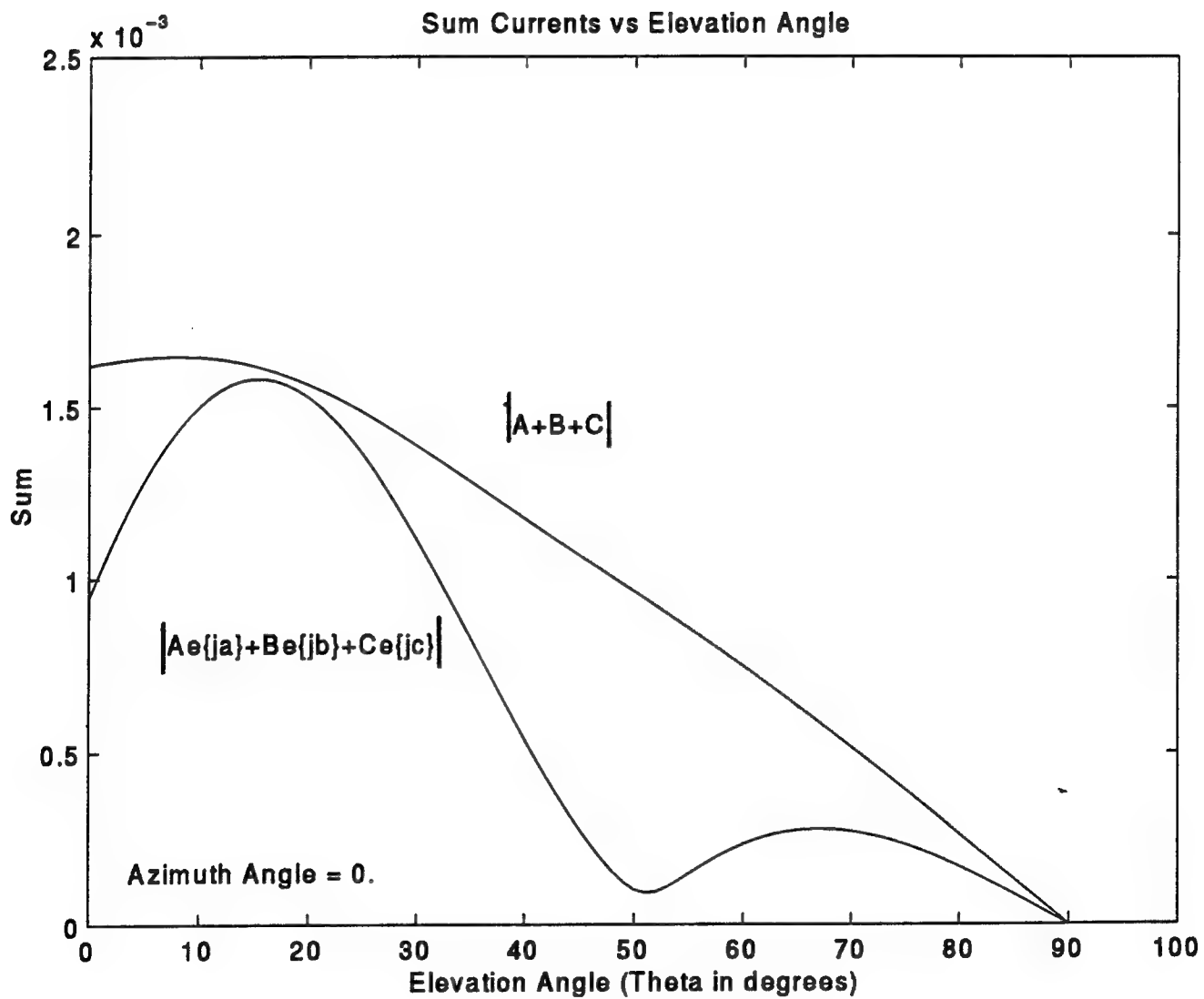


Figure 2

### Layout of Reference Antenna Model

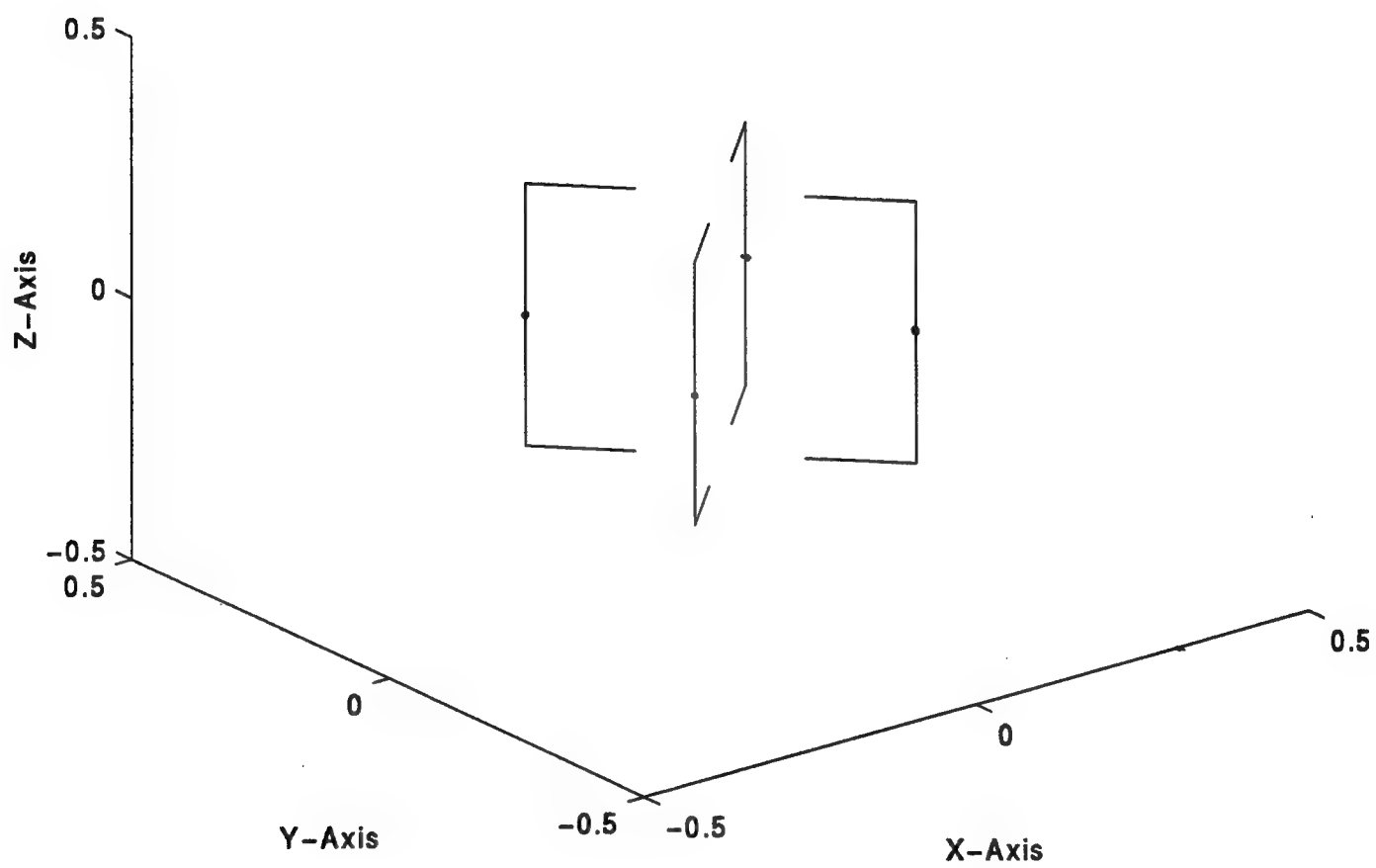


Figure 3

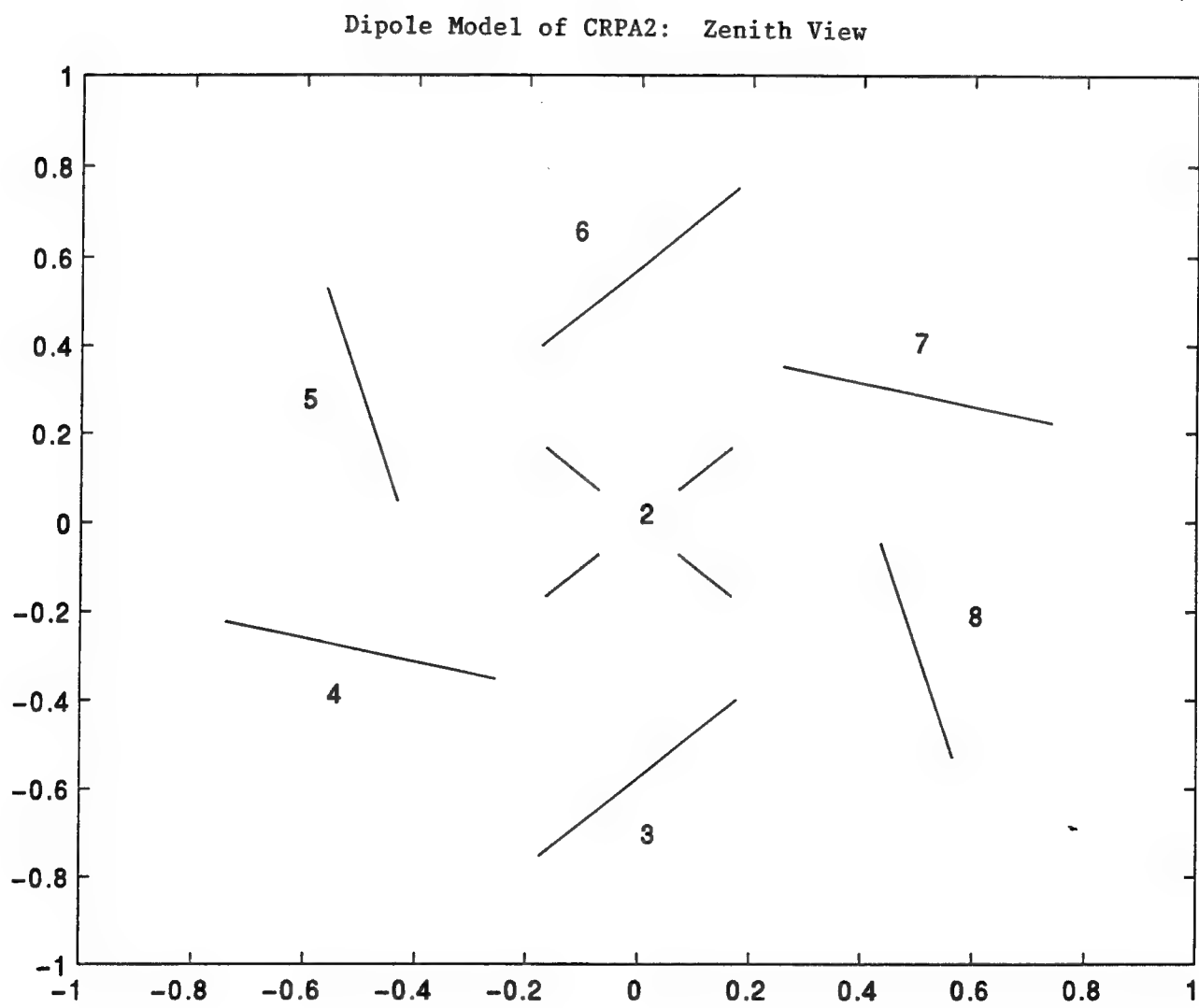
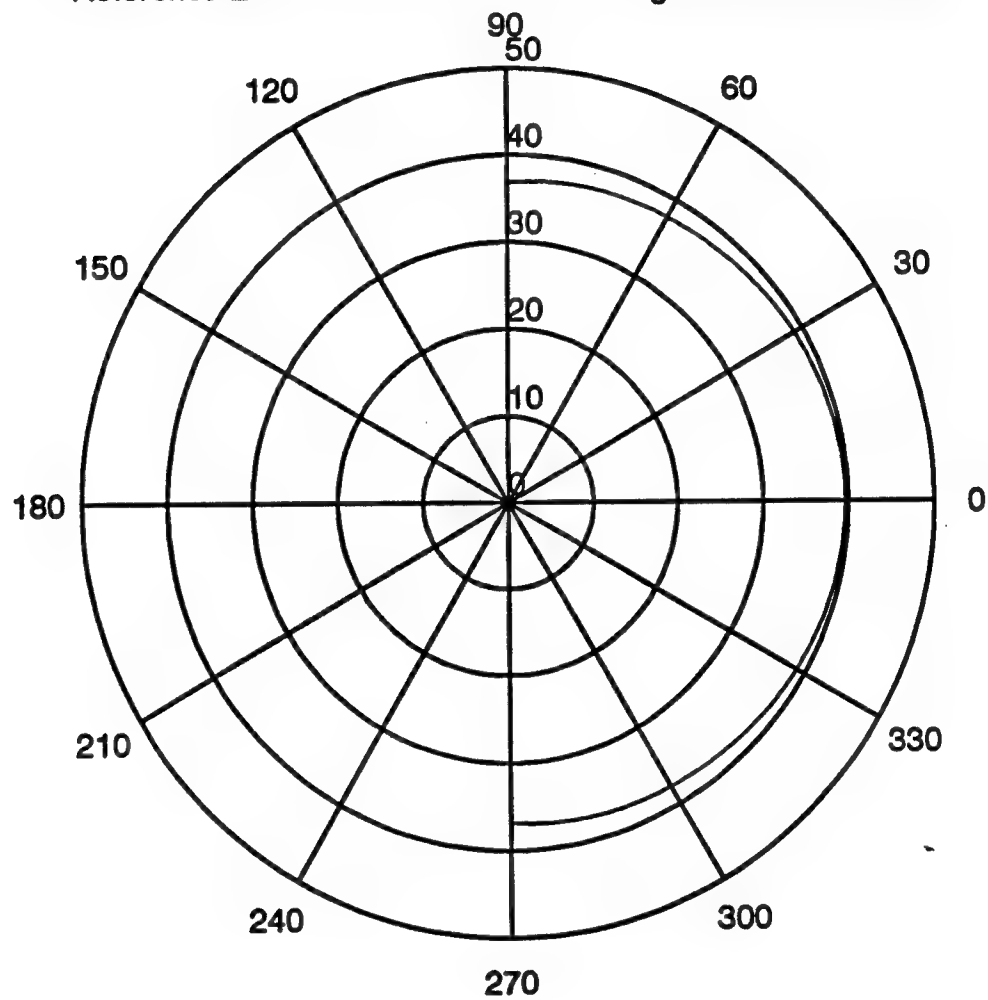


Figure 4

Reference Element Terminal Current Magnitude vs Elevation



Azimuth = 90  
(Calculated)

Figure 5



# ISSUES IN INITIATION AND PROPAGATION OF DETONATION IN REACTIVE SOLIDS

Joseph M. Powers  
Associate Professor  
Department of Aerospace and Mechanical Engineering

University of Notre Dame  
Notre Dame, Indiana 46556-5637

Final Report for:  
Summer Faculty Research Program  
Wright Laboratory  
Eglin Air Force Base, Florida

Sponsored by:  
Air Force Office of Scientific Research  
Bolling Air Force Base, DC

and

Wright Laboratory  
Eglin Air Force Base, Florida

November 1995

## TOPICS IN INITIATION AND PROPAGATION OF DETONATION IN REACTIVE SOLIDS

Joseph M. Powers  
Associate Professor  
Department of Aerospace and Mechanical Engineering  
University of Notre Dame

### Abstract

Issues in the transient development of detonations in reactive solid materials have been studied. A macroscopic two-phase model was employed to predict the shock-to-detonation (SDT) in a granulated solid explosive. The long time results were shown to be fully resolved and in agreement with independent predictions from steady theory. Additionally, prediction of the transition time to detonation agreed well with other unsteady theories. In a second effort, a simple microscopic model of reactive shear bands was developed in order to focus on the small scale mechanisms of initiation. It is planned to solve these equations and compare results with those obtained in a coupled experimental study employing a torsional Hopkinson bar.

# ISSUES IN INITIATION AND PROPAGATION OF DETONATION IN REACTIVE SOLIDS

Joseph M. Powers

## 1 Introduction

### 1.1 Combustion overview

Combustion, which can be thought of as a strongly exothermic chemical reaction process, is of great practical importance and at the same time scientifically challenging. Combustion in gaseous systems is based on the mechanism of molecular collisions, and as such is ultimately diffusion controlled at the microscale level. In typical gas phase combustion, molecules of fuel mix, both convectively and diffusively, with molecules of oxidizer. Fuel-oxidizer collisions occur, and for those collisions of sufficient strength, chemical bonds are broken, converting the chemical energy into kinetic energy, which in turn tends to make future collisions more likely to be of sufficient strength to break more chemical bonds. Thus, the process has a tendency to accelerate. On a macroscale, the increase in average kinetic energy is manifested in an increase in temperature.

In solid systems, convection and mixing is typically not nearly as dominant a mechanism. Typically, fuel and oxidizer are finely ground and combined in a heterogeneous slurry. Sensitivity to ignition becomes a strong function of the fineness to which both were ground. Even more sensitive are those materials which effectively have fuel and oxidizer bound into the same large molecular structure.

Both combustible gases and solids admit propagating combustion waves. The types of waves can be broadly categorized as deflagrations or detonations. Relative to detonations, deflagration waves travel at lower speeds and give rise to lower pressures. Typical gas phase deflagration and detonation pressures are  $10^0 \text{ atm}$  and  $10^2 \text{ atm}$ , respectively. Roughly speaking, because their typical initial density is three orders of magnitude larger than typical gas densities, common deflagration and detonation pressures in energetic solids are  $10^3 \text{ atm}$  and  $10^5 \text{ atm}$ , respectively.

### 1.2 Practical motivations

In the studies undertaken during the previous summer, the focus was on combustion in energetic solids. There are many practical motivations for such studies which have direct relevance to the Air Force. These include:

- Suppression of sympathetic detonation in munition storage—sympathetic detonation is the process which occurs when the detonation of a single unit of munition (e.g. one bomb) causes other units to detonate. In munition storage facilities, it is of great importance to have accurate criteria for minimum separation distances for safe storage.
- Prevention of accidental transition to detonation in propellants—It is important to quantify the level of stimuli necessary for detonation initiation so as to have good design criteria for explosive handling facilities.
- Precise control of initiation in hard target penetration weapons—In such weapons, which must withstand severe shock loading without initiation of reaction, an accurate characterization of the explosive is critical.
- Characterization of initial transients in timing devices—It is well-known that detonation wave speeds ( $\sim 7,000 \frac{\text{m}}{\text{s}}$ ) in solids vary as a function of initial bulk density. A precise knowledge of these speeds would make possible more accurate timing devices which rely on propagation of combustion waves to trigger other events.
- Establishment of precise manufacturing criteria—A full understanding of the combustion process would allow the establishment of much more precise criteria for the manufacture of energetic materials, so as to have explosives tailor-made for particular tasks.

### 1.3 Mechanisms for initiation in energetic solids

Probably the most important property of an energetic material is its global heat of reaction. Knowledge of this property tells the designer the magnitude of useful work which may be performed by the combustion process. Determination of this property can be done using well-developed time-independent techniques such as calorimetry.

Nearly as important and much more difficult, both conceptually and experimentally, is a knowledge of the transient behavior. Initiation in general demands an energy stimulus; it is well-known that the ultimate destiny of the energetic material is a strong function both of the magnitude of the stimulus and the rate at which it is applied. For example, a fixed amount of thermal energy, deposited slowly in the material, will diffuse out of a system faster than energy is generated by reaction to sustain the reaction. The same amount of energy localized in time and space, will ignite a section of the material, and cause the reaction to spread throughout the material. In another example, a fixed amount of mechanical energy applied slowly will propagate in the form of nearly isentropic elastic waves. The higher temperatures generated by the dissipation of this energy into thermal energy will be diffused out before reaction can commence. The same stimulus, localized in time and space, will involve much more plastic, dissipative behavior, and give rise to local high temperature regions, known as hot spots. Near hot spots, significant reaction will ensue and accelerate to ignite the entire material.

In a broad sense one might imagine the development of hot-spots in the context of a cascading of length scales. Initially, one might input a mechanical energy pulse distributed over a macroscopic scale. As this energy pulse propagates through the material, geometric complexities, non-linear wave phenomena, and material heterogeneities, will induce a breakup into fine scale structures. For example, in the non-linear wave mechanics present in the problem, interference of two simple waves is likely to spawn a multiplicity of wave structures which will go on to spawn even more. In the interference process itself there can be amplification of relevant properties such as temperature, which are important in initiation.

A detailed understanding of this localization process is presently lacking. A variety of microscale mechanisms have been proposed to explain the formation of hot spots. These include:

- Shock-void interactions—when a shock wave encounters a local void, there is a diffraction process which can generate local regions of high deformation, plastic work, with a consequent temperature increase
- Shear band formation—high strain rate conditions can induce a loss of yield stress due to thermal softening accompanied by an subsequent increase in temperature due to increased viscous dissipation; such an event is highly localized and limited by diffusion over very small length scales.
- Friction—particle-particle rubbing contacts will induce localized frictional heating which could induce reaction.
- Void collapse—when a solid is compressed, gas-containing embedded voids will be squeezed, inducing an adiabatic heating of the entrapped gas which in turn could induce local combustion near the void.

## 2 Common reaction rate models

Generally the key ingredient in a model used to predict initiation is the reaction rate (or kinetic) model. Relative to solids, gas phase kinetic models are well developed. Such is possible because of the broader range of diagnostic techniques available for gases. For example, optical spectroscopic techniques allow time and spatially resolved measurements of individual species concentrations under a variety of operating conditions. Consequently one is able to develop models which predict these phenomena. Typically for an individual reaction step, one might find the rate well-modelled by a

so-called Arrhenius law:

$$\frac{d\lambda}{dt} = kT^\alpha P^\beta \exp(-E/RT)$$

where  $\lambda$  is the dimensionless reaction progress variable,  $k$  is a rate constant,  $T$  is the temperature,  $P$  is the pressure,  $E$  is the activation energy,  $R$  is the gas constant, and  $\alpha$  and  $\beta$  are constants.

Because one does not have the equivalent ability to see inside of a solid, measurements of detailed kinetic data are much more difficult in these materials. A common macroscale model is given by

$$\frac{dr}{dt} = aP^n$$

where  $r$  is the local particle radius,  $a$  is a constant,  $P$  is the local pressure of the gas phase products, and  $n$  is an experimentally determined constant. Such a model is obtained under steady state combustion conditions and allows for no microstructural effects. Other common models with similar deficiencies are the so-called Forest Fire and Arrhenius models.

### 3 Focus of summer 1995 efforts

The author's efforts during the summer were focused on four areas:

- Familiarization with problems and experimental techniques in studying initiation of solid explosives—In this, the author had little experience and the many interactions with personnel at the AWEF at Eglin AFB were invaluable in helping him understand the state of the art.
- Review of continuum mechanics of solids—As the author's main research area has been in reactive fluid mechanics, it was necessary to do much reading and hold discussions concerning many basic issues in high strain rate solid mechanics. In this the author learned much from AWEF personnel and Prof. James J. Mason, also visiting from the University of Notre Dame.
- Hydrocode modeling of transition to detonation—A significant effort was made in continuing an ongoing study of the transition to detonation in granular solid explosives. This work was done in conjunction with the author's graduate student and co-visitor to Eglin AFB, Mr. Keith A. Gonthier.
- Modeling torsional Hopkinson bar—in conjunction the author's graduate student and co-visitor to Eglin AFB, Mr. Richard J. Caspar, along with Prof. Mason, the author began a modeling effort focused on simulating a reactive shear band in a torsional Hopkinson bar.

The last two points will be discussed in some more detail next.

#### 3.1 Hydrocode modeling of shock-to-detonation transition (SDT) in solid explosives

Motivated by the problem of predicting detonations in granulated energetic material, a good portion of the author's work at Eglin was devoted to working with a relevant set of model equations to predict such events. This is part of a long-term effort in which the author has been involved over the past ten years. With full details given in the paper Powers and Gonthier<sup>1</sup>, the aim here was to refine a predictive capability for the intermediate events leading to detonation. That is to say, the model itself does not directly address microstructural details; it does address the details of the acceleration process of an existing low speed combustion process to a high speed detonation.

Our predictions show that in response to a piston impact a leading compaction wave propagates into the granular mixture. The passage of such a wave leaves the material in an essentially voidless

<sup>1</sup>Powers, J. M., and Gonthier, K. A., "A numerical investigation of transient detonation in granulated material," 15th International Colloquium on the Dynamics of Explosions and Reactive Systems, August 1995, Boulder, Colorado, submitted to *Shock Waves*.

state. Our present kinetics model predicts that significant combustion is initiated later in time at the piston face. Consequently, a high speed detonation wave is generated which overtakes the lead compaction wave. In the far field, the wave relaxes to a shock followed by a zone of chemical reaction. The predictions of the unsteady code showed excellent agreement with those of an independent steady state theory. Further, it was possible to well correlate detonation wave speeds, pressures, and transition zone lengths with experimental results.

### 3.2 Modeling of reactive shear banding in torsional Hopkinson bar tests

A second significant effort was made at the early stages of development of a model for reactive shear banding. We adopted many elements of a simple model used by many authors and proposed to subject this model to a new battery of standard analytical techniques to better elucidate its behavior. Such may be necessary as it is estimated that shear band thicknesses can exist on sub micron length scales. Such scales are essentially impossible to model simultaneously with the centimeter or greater macro scales common in multi-dimensional explosive systems.

#### 3.2.1 Assumptions

The following model assumptions are adopted:

- thin shelled cylindrical geometry
- incompressible material
- viscoplastic material
- $\mathbf{v} = (0, v_\theta, 0)$
- $\nabla = (0, 0, \frac{\partial}{\partial z})$
- perfectly plastic yielding
- temperature-dependent yield stress
- constant viscosity
- constant conductivity
- Arrhenius reaction rate model
- velocity and temperature boundary conditions
- initial velocity and temperature profile

#### 3.2.2 Dimensional equations

We suggest the following model equations:

$$\rho \frac{\partial v_\theta}{\partial t} = \frac{\partial \sigma_{z\theta}}{\partial z} \quad (1)$$

$$\rho \frac{\partial e}{\partial t} = \sigma_{z\theta} \frac{\partial v_\theta}{\partial z} - \frac{\partial q_z}{\partial z} \quad (2)$$

$$\frac{\partial \lambda}{\partial t} = A(1 - \lambda) \exp\left(-\frac{E}{RT}\right) \quad (3)$$

$$\sigma_{z\theta} = \sigma_y(T) \operatorname{sgn}\left(\frac{\partial v_\theta}{\partial z}\right) + \mu \frac{\partial v_\theta}{\partial z} \quad (4)$$

$$e = cT - \lambda Q \quad (5)$$

$$q_z = -k \frac{\partial T}{\partial z} \quad (6)$$

$$\sigma_y(T) = \hat{\sigma} \exp\left(-\frac{T}{T_m}\right) \quad (7)$$

Equation 1 represents conservation of linear momentum; Eq. 2, conservation of energy. Eq. 3 gives a simple one-step kinetic law. Eq. 4 is the viscoplastic relationship for stress. Eq. 5 is the caloric state equation. Eq. 6 is Fourier's law of heat conduction, and Eq. 7 is a postulated equation for the behavior of yield stress with temperature. Here  $\rho$  is the density;  $v_\theta$  is the velocity in the direction of rotation;  $\sigma_{z\theta}$  is the hoop stress;  $e$  is the internal energy;  $q_z$  is the longitudinal heat flux;  $\lambda$  is the reaction progress;  $T$  is the temperature;  $z$  is the longitudinal distance, and  $t$  is time. Constants are a rate constant  $A$ , the activation energy  $E$ , the gas constant  $R$ , the viscosity  $\mu$ , the specific heat  $c$ , the heat of reaction  $Q$ , the thermal conductivity  $k$ , a reference stress  $\hat{\sigma}$ , and a transition temperature  $T_m$ .

Eliminating the algebraic equations and imposing a set of initial and boundary conditions, the equations can be simplified to

$$\rho \frac{\partial v_\theta}{\partial t} = \frac{\partial}{\partial z} \left( \hat{\sigma} \exp \left( -\frac{T}{T_m} \right) + \mu \frac{\partial v_\theta}{\partial z} \right) \quad (8)$$

$$\begin{aligned} \rho c \frac{\partial T}{\partial t} = & \left( \hat{\sigma} \exp \left( -\frac{T}{T_m} \right) + \mu \frac{\partial v_\theta}{\partial z} \right) \frac{\partial v_\theta}{\partial z} \\ & + \rho A Q \exp \left( -\frac{E}{RT} \right) + k \frac{\partial^2 T}{\partial z^2} \end{aligned} \quad (9)$$

$$\frac{\partial \lambda}{\partial t} = A (1 - \lambda) \exp \left( -\frac{E}{RT} \right) \quad (10)$$

$$\lambda(z, 0) = 0 \quad (11)$$

$$v_\theta(z, 0) = V(z), \quad T(z, 0) = T_o \quad (12)$$

$$v_\theta(0, t) = V(0) = V_o, \quad T(0, t) = T_o \quad (13)$$

$$v_\theta(L, t) = V(L) = 0, \quad T(L, t) = T_o \quad (14)$$

### 3.2.3 Dimensionless equations

We scale all variables by known constants:

$$z^* = \frac{z}{L}, \quad t^* = \frac{V_o t}{L} \quad (15)$$

$$(16)$$

$$v^* = \frac{v_\theta}{V_o}, \quad T^* = \frac{T - T_o}{T_o}, \quad \lambda^* = \lambda \quad (17)$$

The following dimensionless parameters arise:

$$Re = \frac{\rho V_o L}{\mu}, \quad Pr = \frac{\mu c}{k}$$

$$Ec = \frac{V_o^2}{c T_o}, \quad \Theta = \frac{E}{R T_o}$$

$$\sigma^* = \frac{\hat{\sigma}}{\rho V_o^2}, \quad T_{mo} = \frac{T_m}{T_o}$$

$$A^* = \frac{L A}{V_o}, \quad Q^* = \frac{Q}{c T_o}$$

We get the following dimensionless model equations, initial and boundary conditions:

$$\begin{aligned} \frac{\partial v^*}{\partial t^*} = & \frac{\partial}{\partial z^*} \left( \sigma^* \exp \left( -\frac{T^* + 1}{T_{mo}} \right) + \frac{1}{Re} \frac{\partial v^*}{\partial z^*} \right) \\ \frac{\partial T^*}{\partial t^*} = & \left( \sigma^* Ec \exp \left( -\frac{T^* + 1}{T_{mo}} \right) + \frac{Ec}{Re} \frac{\partial v^*}{\partial z^*} \right) \frac{\partial v^*}{\partial z^*} \end{aligned} \quad (18)$$

$$+A^*Q^*\exp\left(-\frac{\Theta}{T^*+1}\right)+\frac{1}{PrRe}\frac{\partial^2 T^*}{\partial z^{*2}} \quad (19)$$

$$\frac{\partial \lambda^*}{\partial t^*} = A^*(1-\lambda)\exp\left(-\frac{\Theta}{T^*+1}\right) \quad (20)$$

$$\lambda^*(z^*, 0) = 0 \quad (21)$$

$$v^*(z^*, 0) = V^*(z^*), \quad T^*(z^*, 0) = 0 \quad (22)$$

$$v^*(0, t) = V^*(0) = 1, \quad T^*(0, t^*) = 0 \quad (23)$$

$$v^*(1, t) = V^*(1) = 0, \quad T^*(1, t^*) = 0 \quad (24)$$

### 3.2.4 Solution techniques

A variety of analytic techniques can be brought to bear on these equations. While ultimately one wants to have the general applicability available in modern finite element codes, for problems such as these with fine scale structures, it is essential to use more refined techniques to capture all of the relevant phenomena. A sampling is given below:

- steady state limit—here we enforce that there are no transients. This represents the long time limit of the equations. Here the partial differential equations reduce to a coupled set of ordinary differential equations, which in this case form a general two-point boundary value problem. Such problems can be solved straightforwardly by a variety of efficient numerical methods. In particular, if a steady shear band event occurs, the method will predict it. Additionally it may be possible to identify non-unique solutions.
- linear stability—with knowledge of the steady solutions, the transient equations can be posed in linear form using each steady state as a base state. In this case, it is likely that a separation of variables technique will be sufficient to solve the linearized transient problem. What will arise is an eigenvalue problem with the eigenvalues and eigenfunctions determined in a simple numerical procedure. In general for each steady solution one will obtain a family of eigenvalues and eigenfunctions. Each negative eigenvalue will correspond to a decaying mode in time. Further the eigenvalues will fix the time scales for relaxation. Any positive eigenvalue will indicate a growing mode and render the solution unstable.
- Asymptotic transient analysis—An asymptotic analysis of the induction phase of reaction should be possible to determine a time for thermal explosion, that is an ignition time. Use of the method of multiple scales could in principle reveal details of the actual transition. Such information would be valuable in the proper time scaling for full numerical simulations.
- Full transient analysis—a full transient analysis solving the full partial differential equations using the method of lines would then expose the full range of solutions from beginning to end in time and space. Such solutions could be checked at every step with the limiting cases developed earlier.
- comparison with finite element solutions—solutions obtained with these methods could then be used in the detailed verification of predictions of finite element models. Problems such as shear band formation currently stretch the limits of existing codes and computers when operated in a general multidimensional time-dependent mode. When properly verified both with rationally obtained limiting cases and with experiments, one will have a greater confidence in these general predictive tools.



## 4 Summary

In summary, the author views the program a success in that he was exposed to a variety of new and challenging problems. In brief, on finds that

- solid initiation problems are difficult due to widely disparate space and time scales present
- it is currently impossible to simultaneously model macroscale phenomena while including detailed microscale initiation events
- coupled experimental/microscale modeling/macroscale modeling effort necessary for achievement of goals

## 5 Acknowledgments

For Dr. Joseph C. Foster, Jr.'s global insight and encouragement, for Prof. James J. Mason's guidance in the fundamentals of solid mechanics, for Mr. Keith A. Gonthier's diligence in calculations of two phase detonations, for Mr. Richard J. Caspar's efforts in reactive shear band modeling and torsional Hopkinson bar tests, and for Mr. David R. Wagnon's generous assistance in trying circumstances, the author is grateful.

# **A STUDY OF THE NEW NEAR-OPTIMAL NONLINEAR CONTROL DESIGN TECHNIQUE**

**STATE DEPENDENT ALGEBRAIC RICCATI EQUATION (SDARE) METHOD**

Zhihua Qu

Associate Professor

Department of Electrical Engineering

University of Central Florida

Orlando, FL 32816, USA

Final report for

Air Force Summer Faculty Program

Wright Laboratory, Eglin AFB, FL 32542

Sponsored by

Air Force Office of Scientific Research

Bolling AFB, Washington DC

and

Wright Laboratory, Eglin AFB, FL 32542

August 4, 1995

# A STUDY OF THE NEW NEAR-OPTIMAL NONLINEAR CONTROL DESIGN TECHNIQUE

## STATE DEPENDENT ALGEBRAIC RICCATI EQUATION (SDARE) METHOD

Zhihua Qu

Associate Professor

Department of Electrical and Computer Engineering

University of Central Florida

Orlando, FL 32816, USA

### Abstract

In this report, global stability of a nonlinear regulator design is studied. The technique under investigation is the state-dependent algebraic Riccati equation technique recently proposed in [4, 5]. A near optimal strategy is proposed under which global asymptotic stability of nonlinear system is guaranteed under point-wise stabilizability of its linearized version and two-point boundary value problem is avoided. The study is done using Lyapunov direct method and is based on judicious choices of weighting matrices  $Q$  and  $R$ . Although the resulting control is only near optimal, the proposed control design method is not only efficient in implementation but also applicable to general nonlinear systems with guaranteed stability and performance.

# A STUDY OF THE NEW NEAR-OPTIMAL NONLINEAR CONTROL DESIGN TECHNIQUE

## STATE DEPENDENT ALGEBRAIC RICCATI EQUATION (SDARE) METHOD

Zhihua Qu

### Abstract

In this report, global stability of a nonlinear regulator design is studied. The technique under investigation is the state-dependent algebraic Riccati equation technique recently proposed in [4, 5]. A near optimal strategy is proposed under which global asymptotic stability of nonlinear system is guaranteed under pointwise stabilizability of its linearized version and two-point boundary value problem is avoided. The study is done using Lyapunov direct method and is based on judicious choices of weighting matrices  $Q$  and  $R$ . Although the resulting control is only near optimal, the proposed control design method is not only efficient in implementation but also applicable to general nonlinear systems with guaranteed stability and performance.

**Key words:** Nonlinear systems, Lyapunov direct method, regulator, optimal control, stability, performance index, stabilizability.

### 1 Introduction

The main difficulty in nonlinear analysis and control design is the fact that nonlinear systems in general have different behaviors depending on their initial conditions and inputs. In order to design an effective control that achieves stability and performance, one must develop analytical means to study property of nonlinear systems without solving explicitly their solutions. The approach that is applicable to general nonlinear systems is the Lyapunov direct method. Generally speaking, the key to a successful application of Lyapunov direct method is to find a Lyapunov function and its associated control that are "adequate" to the system dynamics under study. In the last few years, progresses have been reported in the areas of nonlinear control designs based on the Lyapunov direct method. The method that has attracted much attention is the so-called recursive design approach in which nonlinear systems is analyzed equation by equation in order to find proper Lyapunov functions and controls. It has been shown in [8, 13] that systems satisfying the cascaded structure can be stabilized as a whole by adaptive control and robust control designed recursively and backwards. Recursive design can also be applied in a forward fashion for the so-called feedforward systems [9]. Most recently, the recursive-interlacing design procedure has been proposed [11, 12] in which dynamics are exploited by both backward and forward interlacing steps so that stabilizability of nonlinear

systems can be studied and nonlinear controls can be generated without imposing any structural conditions.

Nonlinear control has also been studied along the line of optimal control theory [1]. For nonlinear systems, nonlinear regulator can be developed by solving Riccati-like partial differential equation [3]. Although this method produces optimal nonlinear control, the existence of solution to partial differential equation and its connection to stabilizability is unclear. In addition, the technique requires that the solution satisfy two-point boundary values. In contrast, a new technique is proposed [5] which, based on nonlinear parameterization, is based on algebraic Riccati equation and requires only finite time solution to one-point boundary value differential equation. Literature survey on available results in this direction can also be found in [5].

The main goal of the summer research is to study global stability of nonlinear systems under the proposed state dependent algebraic Riccati equation method. It will be shown in this report that global asymptotic stability can be guaranteed by judiciously choosing a scalar multiplier function in weighting matrices  $Q$  and  $R$  and by solving a proper one-point boundary-value problem.

## 2 Nonlinear Optimal Control

Consider the following nonlinear, affine system

$$\dot{x} = f(x) + g(x)u, \quad (1)$$

where  $x \in \mathbb{R}^n$  and  $u \in \mathbb{R}^m$ , and functions  $f(\cdot)$  and  $g(\cdot)$  are continuous. The objective is to devise a nonlinear, continuous control

$$u = \phi(x), \quad (2)$$

such that the closed-loop, autonomous system

$$\dot{x} = f(x) + g(x)\phi(x) \quad (3)$$

is globally asymptotically stable. This stabilization problem can be formulated into an optimal control problem as follows.

Define the performance index to be

$$J(x(t_0), u, t_0, t_f) = \frac{1}{2}x^T(t_f)Sx(t_f) + \frac{1}{2} \int_{t_0}^{t_f} (x^T Q x + u^T R u) dt, \quad (4)$$

where  $t_f \in [t_0, \infty]$  is the time interval of optimization, and  $S$  is a given constant positive definite matrix. Matrices  $Q$  and  $R$  can be functions of  $x$ . For simplicity, we shall limit our attention to the case that

$$Q(x) = e^{q(x)} Q_0, \quad R(x) = e^{r(x)} R_0, \quad (5)$$

for some constant positive definite matrices  $Q_0$  and  $R_0$  and for some scalar, non-negative functions  $q(x)$  and  $r(x)$ . The optimal control problem is to find the optimal control  $u^*$  that minimizes the performance index,

that is,

$$J(x(t_0), u, t_0, t_f) \geq J^* \triangleq J(x(t_0), u^*, t_0, t_f) \quad \forall u \in \mathbb{R}^m.$$

It is well known [6] that a differential equation whose right hand side contains only continuous functions has a uniformly continuous solution over finite interval and that, if the solution stays in a finite region (that is, a stable trajectory), the solution is unique and has maximum continuation over infinite time interval. By Babarlat lemma [14], we know that, if a control in the form of (2) minimizes the performance index and if the minimum is finite over infinite time horizon, the resulting closed loop system (3) is asymptotically stable. This conclusion can be restated as the following fundamental lemma.

**Lemma 1:** System (1) is globally asymptotically stable under control (2) if  $u = u^*$  and if

$$J^* = J(x(t_0), u^*, t_0, \infty) < \infty.$$

In what follows, the solution to optimal control problem will be provided and analyzed.

## 2.1 Optimality Conditions

Consider system (1) and define a state-dependent parameterization as

$$f(x) = A(x)x$$

where matrix  $A(x)$  is assumed to be well defined for all  $x \in \mathbb{R}^n$ . Then, system (1) can be rewritten as

$$\dot{x} = A(x)x + B(x)u, \quad (6)$$

where  $B(x) = g(x)$ .

The necessary conditions for optimality can be found using the calculus of variations. To this end, we form the Hamiltonian  $H$  as

$$H = \frac{1}{2}x^T Qx + \frac{1}{2}u^T Ru + \lambda^T [f(x) + B(x)u], \quad (7)$$

where  $\lambda \in \mathbb{R}^n$  is the Lagrangian multiplier. Then, the necessary conditions for optimality are [2]:

$$\dot{x} = \frac{\partial H}{\partial \lambda}, \quad \frac{\partial H}{\partial u} = 0, \quad \dot{\lambda} = -\frac{\partial H}{\partial x}.$$

Since performance index (4) is convex, any stationary point will be at least a local optimum. Thus, there is no need to check second-order conditions associated with the Hamiltonian.

It follows that the condition  $\dot{x} = \partial H / \partial \lambda$  is always satisfied. It follows from the condition  $\partial H / \partial u$  that

$$\frac{\partial H}{\partial u} = Ru + B^T \lambda.$$

By requiring a optimal control candidate in (2) to be of the form

$$u = -R^{-1}B^T Px \quad (8)$$

for some matrix function  $P(x)$ , the condition  $\partial H \partial u = 0$  can be rewritten as

$$B^T[\lambda - Px] = 0,$$

which can be satisfied if the Lagrangian multiplier is chosen to be

$$\lambda = Px. \quad (9)$$

Note that control (8) is a nonlinear state feedback control and that it becomes the optimal control if matrix  $P(x)$  can be selected to satisfy the third and the last necessary condition. If the optimal control exists, substituting control (8) into (6) yields the optimal closed loop system

$$\dot{x} = (A - BR^{-1}B^TP)x. \quad (10)$$

Having found the multiplier, we can proceed to check the third necessary condition. Note that

$$\begin{aligned} \dot{\lambda} &= \dot{Px} + P\dot{x} \\ &= \dot{Px} + PAx - PBR^{-1}B^TPx, \end{aligned}$$

and that

$$\begin{aligned} \frac{\partial H}{\partial x} &= Qx + \frac{1}{2} \frac{\partial q(x)}{\partial x} x^T Qx + \frac{1}{2} \frac{\partial r(x)}{\partial x} u^T Ru + \left( \frac{\partial f}{\partial x} \right)^T \lambda \\ &\quad + \text{vec} \left\{ u^T \frac{\partial B}{\partial x_i} \lambda \right\} \\ &= Qx + \frac{1}{2} \frac{\partial q(x)}{\partial x} x^T Qx + \frac{1}{2} \frac{\partial r(x)}{\partial x} u^T Ru \\ &\quad + \text{vec} \left\{ x^T PBR^{-1} \frac{\partial B}{\partial x_i} Px \right\} + \left( \frac{\partial f}{\partial x} \right)^T Px, \end{aligned}$$

where

$$\text{vec}\{z_i\} \triangleq \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix}.$$

It follows from  $f(x) = A(x)x$  that

$$\frac{\partial f}{\partial x} = A + \text{vec} \left\{ x^T \left( \frac{\partial A_i^T}{\partial x} \right)^T \right\}, \quad (11)$$

where  $A_i$  is the  $i$ -th row of matrix  $A$ . By direct computation, one can show that

$$\text{vec} \left\{ x^T \left( \frac{\partial A_i^T}{\partial x} \right)^T \right\} = \begin{bmatrix} \frac{\partial A}{\partial x_1} x & \cdots & \frac{\partial A}{\partial x_n} x \end{bmatrix}. \quad (12)$$

Therefore, the third necessary condition of optimality can be rewritten to be

$$\begin{aligned} 0 &= \dot{Px} + (PA + A^TP - PBR^{-1}B^TP + Q)x + \frac{1}{2} \frac{\partial q(x)}{\partial x} x^T Qx \\ &\quad + \frac{1}{2} \frac{\partial r(x)}{\partial x} u^T Ru + \text{vec} \left\{ x^T \frac{\partial A^T}{\partial x_i} Px + x^T PBR^{-1} \frac{\partial B}{\partial x_i} Px \right\}. \end{aligned} \quad (13)$$

Since the first two necessary conditions have been satisfied, equation (13) is the optimality condition.

In summary, the optimal control problem can be solved if solution  $P(x(t_0), t_0, t_f)$  to the nonlinear partial differential equation (13) can be found. Obviously, solving the partial differential equation is not easy; in particular, it is a two-point boundary-value problem satisfying

$$x(t_0) \text{ given, } P(x(t_0), t_0, t_f) = S,$$

and

$$0 < \lim_{t_f \rightarrow \infty} P(x(t_0), t_0, t_f) < \infty. \quad (14)$$

Therefore, the optimal solution can only be found numerically by backward and forward sweeps. In fact, this is the key and also the main difficulty of applying optimal control methodology to nonlinear systems.

An interesting observation that is pivotal to our SDARE method, a suboptimal design, is the following. If matrix  $S$  is not fixed apriori and can be chosen freely, the optimality condition becomes an one-point boundary-value problem and therefore, for any initial condition and for any finite  $t_f$ , it can be solved by numerical integration (*forward sweep only*). If the corresponding optimal control exists and can stabilize the system, such a solution will force the state approach the origin. However, while the state moves toward the origin, we can not implement indefinitely this one-point boundary-value solution because the solution may grow without bound and thus violate the requirement (14). Intuitively, unboundedness of the one-point boundary value solution can be explained as follows. As the state approaches zero, the left hand side of the optimality condition (13) becomes close to zero and yields an equation whose solution, in the scalar case, is of zero over zero. The solution of this type of equations could be anything including, again in the scalar case for example, a finite number, positive infinity, negative infinity. In the case that the solution becomes larger and larger, implementation of the control will become impossible numerically. Whether the limit of the solution is finite depends on its initial condition. The solution has a finite limit is the optimal solution over infinite horizon and, as discussed before, it can only be found by backward and forward sweeps. Nevertheless, we know that the optimality condition can be enforced numerically by forward sweep for a finite amount of time.

Besides the two-point boundary-value problem, the following two problems associated with the optimal control are also essential to its applications. First, assuming that the optimality condition is satisfied, is the closed loop system stable? The stability property under optimal control is a non-trivial problem for nonlinear systems. For nonlinear systems, stability analysis can be done using the Lyapunov second method in which finding a proper Lyapunov function is the key. Lyapunov function and explicit condition for achieving stability will be provided in subsection 2.3. The second question is how to choose matrices  $Q$  and  $R$  so that the closed loop system has better performance. Since stability is always part of performance requirement, the second problem is obviously related to the first. On the other hand, since there is no



general performance measure except exponential convergence rate defined for nonlinear systems, performance improvement through choices of  $Q$  and  $R$  is often done empirically rather than analytically.

The main contributions of this report are two items: (i) Condition is provided for global asymptotic stability of nonlinear systems under an optimal control is established using Lyapunov direct method (section 2.3). (ii) The SDARE method is analyzed and a method of switching on and off the optimality condition is proposed in order to get around the two-point boundary-value problem while maintaining global stability (section 3). Specifically, choices of matrices  $Q$  and  $R$  are found explicitly so that, while enforcing the optimality condition for a finite period of time, the stability conditions mentioned in (i) is met; once the state enters a small neighborhood around the origin, the optimality condition is no longer maintained to avoid the two-point boundary-value problem and to overcome possible numerical problem associated with imposing one-point boundary-value optimality condition indefinitely; and Lyapunov direct method is applied to demonstrate global asymptotic stability of the scheme (section 3.2). This study establishes that the sub-optimal control design method based on state-dependent algebraic Riccati equation is effective and efficient. Its simplicity and wide applicability makes it very useful in many engineering applications.

## 2.2 Simplifications of Optimality Condition

In this subsection, we shall provide several simplified versions of the optimality condition (13) that are useful in simplifying the subsequent analysis and presentation. Obviously, all versions of the optimality condition have the same problems discussed before.

The first simplified version is obtained based on the observation: for any nonlinear system whose matrix  $B(x)$  is rank invariant (locally or globally), there is an invertible state transformation so that, in the transformed state space,

$$B(x) = \begin{bmatrix} 0_{(n-m) \times m} \\ I_{m \times m} \end{bmatrix},$$

where  $I_{m \times m}$  is the  $m$ -dimensional identity matrix and  $0$  is the zero matrix of  $(n - m)$  by  $m$ . Therefore, we can assume without loss of any generality that matrix  $B(x)$  be constant. In the subsequent discussions,  $B$  being constant is always implied unless stated otherwise. In this case, the optimality condition reduces to

$$\begin{aligned} 0 = & \dot{P}x + (PA + A^T P - PBR^{-1}B^T P + Q)x + \frac{1}{2} \frac{\partial q(x)}{\partial x} x^T Q x \\ & + \frac{1}{2} \frac{\partial r(x)}{\partial x} u^T R u + \text{vec} \left\{ x^T \frac{\partial A^T}{\partial x_i} P x \right\}. \end{aligned} \quad (15)$$

In addition to matrix  $B$ , matrices  $Q$  and  $R$  are also subject to partial differentiation in the optimality condition. Since matrices  $Q$  and  $R$  are chosen by the designer, they can be made to be constant. Specifically, optimality condition (15) reduces to

$$\dot{P}x + (PA + A^T P - PBR^{-1}B^T P + Q)x + \frac{1}{2} \frac{\partial q(x)}{\partial x} x^T Q x + \text{vec} \left\{ x^T \frac{\partial A^T}{\partial x_i} P x \right\} = 0, \quad (16)$$

if  $R$  is constant; and to

$$\dot{P}x + (PA + A^T P - PBR^{-1}B^T P + Q)x + \text{vec} \left\{ x^T \frac{\partial A^T}{\partial x_i} \right\} Px = 0, \quad (17)$$

if both  $R$  and  $Q$  are constant. One may write down the optimality condition for the case that  $Q$  is constant but  $R$  is not. However, this case is redundant since the corresponding equation may be transformed into one in which the new triplet  $\{Q, R, P\}$  has the property that  $R$  is constant while  $Q$  is not. As to be shown later in sections 3.2 and 3.3,  $Q(x)$  and  $R(x)$  must be chosen to be functions of  $x$  and the optimality condition (15) should be used for general nonlinear systems; even for scalar nonlinear system,  $Q$  and  $R$  can not be made constant simultaneous since condition (17) can never be satisfied globally.

The last simplification is that, if the system is linear,  $\frac{\partial f}{\partial x} = A$  and  $\frac{\partial A}{\partial x_i} = 0$ . In this case, the optimality condition (17) reduces to the well known differential Riccati equation for linear systems. Obviously, this simplification is included here only for completeness.

## 2.3 Stability of Optimal Control Systems

In this section, stability property of optimal control systems will be analyzed. To this end, we assume that optimality condition (16) be satisfied. Similar results can be derived for other versions of the optimality condition.

It follows from (10), (8), and (15) that

$$\begin{aligned} \frac{d}{dt}(x^T P x) &= x^T \dot{P}x + 2x^T P \dot{x} \\ &= x^T (\dot{P} + PA + A^T P - 2PBR^{-1}B^T P)x \\ &= -x^T (Q + PBR^{-1}B^T P)x - \frac{1}{2}x^T \frac{\partial q(x)}{\partial x} x^T Qx \\ &\quad - \frac{1}{2}x^T \frac{\partial r(x)}{\partial x} u^T Ru - x^T \text{vec} \left\{ x^T \frac{\partial A^T}{\partial x_i} \right\} Px. \end{aligned} \quad (18)$$

Using the above relationship, we can rewrite the performance index as

$$\begin{aligned} J^* &= \frac{1}{2}x^T(t_f)Sx(t_f) + \frac{1}{2} \int_{t_0}^{t_f} (x^T Qx + u^T Ru)dt \\ &= \frac{1}{2}x^T(t_f)Sx(t_f) - \frac{1}{2} \int_{t_0}^{t_f} d(x^T P x) - \frac{1}{4} \int_{t_0}^{t_f} \left[ x^T \frac{\partial q(x)}{\partial x} x^T Qx \right. \\ &\quad \left. + x^T \frac{\partial r(x)}{\partial x} u^T Ru + 2x^T \text{vec} \left\{ x^T \frac{\partial A^T}{\partial x_i} \right\} Px \right] dt. \end{aligned}$$

Therefore, we have that, if boundary condition  $P(x(t_0), t_0, t_f) = S$  is satisfied,

$$\begin{aligned} J^* &= \frac{1}{2}x^T(t_0)P(x(t_0), t_0, t_0)x(t_0) - \frac{1}{4} \int_{t_0}^{t_f} \left[ x^T \frac{\partial q(x)}{\partial x} x^T Qx \right. \\ &\quad \left. + x^T \frac{\partial r(x)}{\partial x} u^T Ru + 2x^T \text{vec} \left\{ x^T \frac{\partial A^T}{\partial x_i} \right\} Px \right] dt, \end{aligned} \quad (19)$$

and that, if matrix  $P(x(t_0), t_0, t_f)$  is not  $S$  but is positive semi-definite,

$$J^* = \frac{1}{2}x^T(t_0)P(x(t_0), t_0, t_0)x(t_0) + \frac{1}{2}x^T(t_f)[S - P(x(t_0), t_0, t_f)]x(t_f) - \frac{1}{4} \int_{t_0}^{t_f} \left[ x^T \frac{\partial q(x)}{\partial x} x^T Q x + x^T \frac{\partial r(x)}{\partial x} u^T R u + 2x^T \text{vec} \left\{ x^T \frac{\partial A^T}{\partial x_i} \right\} P x \right] dt. \quad (20)$$

It is worth noting that  $J^*$  being finite as  $t_f$  approaches infinity can not be concluded in general from either (19) or (20) without any condition. This is because, although the optimal control is the best available in the sense that it makes performance index minimum, optimality does not guarantee that system (6) is stabilizable or that control (8) is stabilizing. Only in the case that the system is linear and stabilizable (stabilizability or, more sufficiently, controllability ensures that solution  $P$  to the algebraic Riccati equation exists),  $J^*$  is obviously finite by choosing a constant, positive definite matrix  $Q$  and therefore global stability can be concluded. Note that, in the case that  $Q$  is chosen to be positive semi-definite for output regulation, detectability (or more sufficiently observability) must also be imposed together with stabilizability to guarantee stability.

Since the system is nonlinear, we adopt for stability analysis the Lyapunov direct method. Consider the Lyapunov candidate:

$$V(x) = x^T P(x)x. \quad (21)$$

Possible choice of other Lyapunov functions will be studied in section 3.3. Using Lyapunov function (21) and recalling (18), we can conclude the following result under optimality condition.

**Lemma 2:** Under optimality condition (16), system (6) under optimal control (8) is globally asymptotically stable if matrix  $P(x)$  is positive definitely not only pointwise everywhere but also uniformly in the sense that, for some constant  $c_0 > 0$  and for all  $x$ ,

$$\lambda_{\min}(P) \geq c_0 > 0, \quad (22)$$

and if the matrix sum

$$W(x) \triangleq Q + PBR^{-1}B^T P + \frac{1}{2}x^T \frac{\partial q(x)}{\partial x} Q + \frac{1}{2}x^T \frac{\partial r(x)}{\partial x} PBR^{-1}B^T P + \frac{1}{2} \text{vec} \left\{ x^T \frac{\partial A^T}{\partial x_i} P + P \frac{\partial A}{\partial x_i} x \right\} \quad (23)$$

is positive definite.

It is worth comparing Lemmas 1 and 3. It follows from (19) and (20) that  $J^*$  is finite if matrix sum

$$\frac{1}{2}x^T \frac{\partial q(x)}{\partial x} Q + \frac{1}{2}x^T \frac{\partial r(x)}{\partial x} PBR^{-1}B^T P + \frac{1}{2} \text{vec} \left\{ x^T \frac{\partial A^T}{\partial x_i} P + P \frac{\partial A}{\partial x_i} x \right\}$$

is positive semi-definite. This condition is obviously more conservative than condition (23). To conclude global stability, Lemma 1 requires that  $P$  be at least positive semi-definite (as indicated in (20)) while Lemma 2 requires condition (22).

As to be shown later, we shall explore various possibilities of matrix  $A(x)$  in the nonlinear parameterization  $f(x) = A(x)x$ . Unnecessary complication can be avoided by removing the partial derivatives of  $A$  from the stability condition. It follows from (11) and (12) that global stability condition (23) can be rewritten as

$$\begin{aligned} W(x) = & Q + PBR^{-1}B^TP + \frac{1}{2}x^T \frac{\partial q(x)}{\partial x} Q + \frac{1}{2}x^T \frac{\partial r(x)}{\partial x} PBR^{-1}B^TP \\ & + \frac{1}{2} \left\{ \left( \frac{\partial f}{\partial x} - A \right)^T P + P \left( \frac{\partial f}{\partial x} - A \right) \right\} \\ > & 0. \end{aligned} \quad (24)$$

Lemma 2 is better than Lemma 1 since positive definiteness of matrix  $W(x)$  in (23) can be evaluated. However, it is not known what are necessary and sufficient conditions under which matrix  $W(x)$  can be made positive definite through choices of  $Q$  and  $R$ . In the next section, a sufficient condition and a pair of  $Q$  and  $R$  are given that make matrix  $W$  positive definite.

### 3 State-Dependent Algebraic Riccati Equation Technique

Due to the difficulties associated with solving matrix  $P$  and with ensuring stability condition (24), it is important to find conditions and design procedure so that a sub-optimal, globally stable and numerically efficient control can be designed. A new sub-optimal technique, called SDARE, has been proposed recently in [4, 5]. Essentially, the technique is based on the solution of an algebraic Riccati equation similar to that for linear time invariant systems except that the algebraic Riccati equation is state dependent. In [4, 5], basic steps of the SDARE method are given, and local stability result is established. It is the purpose of this report to develop conditions and suboptimal control strategy under which the SDARE method produces numerically efficient and globally stabilizing controls.

#### 3.1 A Sub-Optimal Solution: SDARE

For the development of the state-dependent algebraic Riccati equation method, we introduce a definition which is extension of that defined for linear time invariant systems.

**Definition:** Matrix  $A(x)$  is a *stabilizable* parameterization for nonlinear system (6) if the pair  $\{A(x), B(x)\}$  is stabilizable pointwise everywhere.

The state-dependent algebraic Riccati equation method consists of three parts. The following is the main design procedure.

##### Main Design Procedure:

- For a given positive integer  $L$ , find off-line  $L + 1$  distinct parameterizations  $A_i(x)$  such that, for  $i = 1, \dots, L + 1$ ,

$$f(x) = A_i(x)x.$$

- For some vector function  $\alpha(x) = \text{vec}\{\alpha_i(x)\} \in \mathbb{R}^L$ , define the composite parameterization  $A(x, \alpha)$  as

$$A(x, \alpha) = \alpha_L A_{L+1}(x) + \sum_{l=1}^L \alpha_{l-1} A_l(x) \prod_{p=l}^L (1 - \alpha_p). \quad (25)$$

- Solve one-line for matrix  $P(x, \alpha)$  from the state-dependent algebraic Riccati equation:

$$A^T(x, \alpha)P(x, \alpha) + P(x, \alpha)A(x, \alpha) - P(x, \alpha)BR^{-1}(x)B^T P(x, \alpha) + Q(x) = 0, \quad (26)$$

where matrices  $R$  and  $Q$  are defined in (5).

- The SDARE nonlinear control law is given by

$$u = -R^{-1}(x)B^T P(x, \alpha)x. \quad (27)$$

It is worth explaining several aspects of the SDARE design procedure. As the main procedure stands,  $\alpha(x)$  can be chosen freely by the designer, and so are matrices  $Q(x)$  and  $R(x)$ . The other two parts of the SDARE design address how to make these choices to guarantee global stability, which is the topic of the next subsection.

The idea of *nonlinear parameterization* was introduced in [5] and its purpose is twofold: (a) The parameterization allows us to introduce functions  $\alpha_i(x)$  that are available for the designer to choose freely. It is the extra degrees of freedom in functions  $\alpha_i(x)$  that allows us to split the optimal condition into two parts, the algebraic Riccati equation and a further simplified optimality condition. (b) The algebraic Riccati equation provides a new avenue to ensure stability conditions (22) and (24) for optimal control system and, at the same time, the new optimality condition can be imposed by a proper choice of  $\alpha(\cdot)$  for a finite period of time without running into the two-point boundary-value problem. Regarding the nonlinear parameterization scheme, we have the following results: one lemma listed below and two conjectures to be provided in the next subsection.

**Lemma 3:** [4, 5] Matrix  $A(x, \alpha)$  in (25) is also a nonlinear parameterization in the sense that  $A(x, \alpha)x = f(x)$ .

Note that, no matter what parameterization is used, function  $A(x, \alpha)x = f(x)$  implies that function  $f(x)$  is independent of  $\alpha$  while  $A(\cdot)$  depends on  $\alpha$ . This is why we rewrite stability condition (23) as (24).

With regard to the algebraic Riccati equation, we have the following results. Lemma 6 can be proven by direct computation.

**Lemma 4:** [4, 5] If parameterizations  $A_i(x)$  ( $i = 1, \dots, L+1$ ) are stabilizable, the parameterization  $A(x, \alpha)$  is also stabilizable, and therefore the solution  $P(x, \alpha)$  to the algebraic Riccati equation (26) is positive definite pointwise everywhere. Furthermore, for any constant  $\alpha$ , the closed loop system under control (27) is locally asymptotically stable around the origin.

**Lemma 5:** [10] If  $\lambda_{\min}(BR_0^{-1}B^T) > 0$  (which can hold only for square systems), the solution of algebraic Riccati equation satisfies the following inequality:

$$\begin{aligned}\lambda_{\max}(P) &\leq \frac{\lambda_{\max}(Q)}{\sqrt{\|A\|^2 + \lambda_{\min}(BR_0^{-1}B^T)\lambda_{\max}(Q)} - \|A\|} \\ &= \frac{e^{q(x)}\lambda_{\max}(Q_0)}{\sqrt{\|A\|^2 + e^{q(x)-r(x)}\lambda_{\min}(BR_0^{-1}B^T)\lambda_{\max}(Q_0)} - \|A\|}.\end{aligned}$$

**Lemma 6:** If  $q(x) = r(x)$ , the solution of algebraic Riccati equation (26) is given by  $P(x, \alpha) = e^{q(x)}P_0(x, \alpha)$  where  $P_0(\cdot)$  is the solution of the new algebraic Riccati equation

$$A^T(x, \alpha)P_0(x, \alpha) + P_0(x, \alpha)A(x, \alpha) - P_0(x, \alpha)BR_0^{-1}B^TP_0(x, \alpha) + Q_0 = 0. \quad (28)$$

The local stability provided by lemma 4 is the foundation of the SDARE method. Intuitively, lemma 4 is a generalization of the result of linear time invariant systems to nonlinear systems. It is expected that such a result can be concluded. The real interesting issue is whether or not global stability and good performance can be guaranteed by control (27). The designer has two choices left: design functions  $\alpha(x)$  and weighting matrices  $Q(x)$  and  $R(x)$ . In the next section, we shall show how to make these choices so that global stability of the closed loop system is guaranteed.

### 3.2 Global Stable SDARE Method

We begin our discussion about how to choose design function  $\alpha(x)$ .

For any given constant vector  $\alpha^\circ$ , expand the closed loop system around the origin using Taylor series as

$$\dot{x} = [A(0, \alpha^\circ) - BR^{-1}(0)BP(0, \alpha^\circ)]x + O(\|x\|^2). \quad (29)$$

Then, there exists a finite neighborhood around the origin,  $\Omega^\circ$ , in which the nonlinear system (29) is locally asymptotically stable. The existence of set  $\Omega^\circ$  is guaranteed by lemma 4. In fact, the local stability proof in [4, 5] was done using the localized Lyapunov function

$$V(x) = x^T P(0, \alpha^\circ)x,$$

which is the lowest order term in the Taylor series expansion of Lyapunov function (21). The fact that both lemmas 2 and 4 use the same Lyapunov function implies that stability conditions in lemma 2 are not needed in the set  $\Omega^\circ$ . This observation will be instrumental in the subsequent discussions. It is well known that finding the maximum size and shape of set  $\Omega^\circ$  is extremely difficult, but for our purpose any conservative estimate is sufficient and such an estimate is not difficult to find [7].

It follows from the algebraic Riccati equation (26) that optimality condition (16) can be satisfied if

$$\dot{P}x + \frac{1}{2} \frac{\partial q(x)}{\partial x} x^T Qx + \frac{1}{2} \frac{\partial r(x)}{\partial x} u^T Qu + \text{vec} \left\{ x^T \frac{\partial A^T}{\partial x_i} Px \right\} = 0.$$

It follows from (11) and (12) that the above equation can be rewritten to be

$$0 = \sum_{i=1}^n \frac{\partial P}{\partial x_i} \dot{x}_i x + \sum_{i=1}^n \frac{\partial P}{\partial \alpha_i} \dot{\alpha}_i x + \frac{1}{2} \frac{\partial q(x)}{\partial x} x^T Q_0 x + \frac{1}{2} \frac{\partial r(x)}{\partial x} u^T Q u + \left( \frac{\partial f}{\partial x} - A \right)^T P x, \quad (30)$$

where  $\dot{x}_i$  can be calculated or substituted by (10). Since algebraic Riccati equation (26) is an identity for all  $\alpha$  and  $x$ , identity remains after taking partial derivative with respect to any scalar variable. That is, we have that, for  $i = 1, \dots, n$ ,

$$[A - BR^{-1}BP]^T \frac{\partial P}{\partial x_i} + \frac{\partial P}{\partial x_i} [A - BR^{-1}BP] = -P \frac{\partial A}{\partial x_i} - \frac{\partial A^T}{\partial x_i} P - \frac{\partial Q}{\partial x_i}, \quad (31)$$

and

$$[A - BR^{-1}BP]^T \frac{\partial P}{\partial \alpha_i} + \frac{\partial P}{\partial \alpha_i} [A - BR^{-1}BP] = -P \frac{\partial A}{\partial \alpha_i} - \frac{\partial A^T}{\partial \alpha_i} P. \quad (32)$$

It is worth noting that  $\frac{\partial A}{\partial x_i}$  and  $\frac{\partial A}{\partial \alpha_i}$  can be computed off-line prior to implementation, that equations (31) and (32) are Lyapunov equations with respect to partial derivatives  $\frac{\partial P}{\partial x_i}$  and  $\frac{\partial P}{\partial \alpha_i}$ , and that, once matrix  $P$  is solved from algebraic Riccati equation (26), the partial derivatives of  $P$  can be solved on-line. After determining the partial derivatives,  $\alpha(x)$  can be solved on-line numerically by integrating the optimality condition (30). That is, we shall use the following procedure to find  $\alpha(x)$  in order to implement control (27) real time.

**Procedure for finding  $\alpha(x)$ :**

- At time  $t = t_i$ , consider the two complementary cases. If  $x \in \Omega^\circ$ , set  $\alpha = \alpha^\circ$ . If  $x \notin \Omega^\circ$ , proceed with the following steps to find  $\alpha$ .
  - First solve  $P$  from algebraic Riccati equation (26).
  - Solve all partial derivatives contained in (30) from (31) and (32).
  - Solve  $\dot{\alpha}$  so that optimality condition (30) holds.
  - Integrate the value of  $\dot{\alpha}$  numerically over the interval  $[t_i, t_{i+1}]$  where  $t_{i+1} = t_i + \delta t$  for sufficiently small  $\delta t > 0$  so that  $\alpha$  at  $t_{i+1}$  is found.
- Repeat the process at  $t = t_{i+1}$ .

Regarding solvability of optimality condition (30) through finding  $\alpha(x)$ , we have the following conjectures about necessary dimension of  $\alpha(x)$ . It is our believe that Conjecture 1 is certainly sufficient since  $n(n-1)/2$  is the number of off-diagonal elements in matrix  $P$  for  $n$ -th order systems. Since the optimality condition consists of  $n$ -th differential equations for  $n$ -th order systems, Conjecture 2 may hold.

**Conjecture 1:** Vector function  $\alpha(x)$  can be found numerically to guarantee equation (30) if index  $L$ , the dimension of  $\alpha(\cdot)$ , is chosen to be  $n(n-1)/2$ .

**Conjecture 2:** Vector function  $\alpha(x)$  can be found numerically to guarantee equation (30) if index  $L$ , the dimension of  $\alpha(\cdot)$ , is chosen to be  $n$ .

It follows from lemma 4 that set  $\Omega^\circ$  is attractive. Therefore, if  $x \in \Omega^\circ$  at time  $t = t_i$ ,  $x \in \Omega^\circ$  for all  $t \geq t_i$  and the state will be asymptotically stable. It also follows from lemma 2 that, by imposing optimality condition, the closed loop system will be globally stable if condition (24) holds. As discussed in section 2, the optimality condition can not be imposed indefinitely without solving off-line the two-point boundary-value problem. Fortunately, optimality condition can always be imposed using only forward sweep for a finite period of time during which, by global stability under condition (24), the system trajectory will enter the set  $\Omega^\circ$ . Therefore, we know that global stability can be achieved without the two-point boundary-value problem if condition (24) holds. Our approach to make matrix  $W(x)$  positive definite is to utilize the algebraic Riccati equation and to pick judiciously matrices  $Q(x)$  and  $R(x)$ . To this end, we have the following lemmas.

**Lemma 7:** Consider the algebraic Riccati equation (26). Then, the matrix sum

$$-(A^T + \Delta A^T)P - P(A + \Delta A) + PBR^{-1}B^T P$$

is positive definite pointwise for any  $\Delta A$  so long as, for all  $x$ ,

$$\|\Delta A\| < \frac{1}{2} \frac{\lambda_{\min}(Q)}{\lambda_{\max}(P)}.$$

Proof: It follows from the algebraic Riccati equation that

$$\begin{aligned} & -(A^T + \Delta A^T)P - P(A + \Delta A) + PBR^{-1}B^T P \\ &= Q - (\Delta A^T P + P \Delta A), \end{aligned}$$

from which the statement of the lemma can be concluded directly.  $\diamond$

**Lemma 8:** Matrix  $W(x)$  can be made positive definite by setting  $r(x) = q(x)$  and by choosing

$$2 + x^T \frac{\partial q(x)}{\partial x} > \frac{1}{2} \left\| \frac{\partial f}{\partial x} - A \right\| \cdot \frac{\lambda_{\min}(Q_0)}{\lambda_{\max}(P_0)}. \quad (33)$$

Proof: It follows from  $r(x) = q(x)$  and from (26) that matrix  $W(x)$  in (24) can be rewritten to be

$$\begin{aligned} W(x) &= \left[ 1 + \frac{1}{2} x^T \frac{\partial q(x)}{\partial x} \right] \cdot (Q + PBR^{-1}B^T P) + \frac{1}{2} \left\{ \left( \frac{\partial f}{\partial x} - A \right)^T P + P \left( \frac{\partial f}{\partial x} - A \right) \right\} \\ &= \left[ 1 + \frac{1}{2} x^T \frac{\partial q(x)}{\partial x} \right] PBR^{-1}B^T P \\ &\quad + \left[ 1 + \frac{1}{2} x^T \frac{\partial q(x)}{\partial x} \right] \cdot [-(A^T + \Delta A^T)P - P(A + \Delta A) + PBR^{-1}B^T P], \end{aligned}$$

where

$$\Delta A = -\frac{1}{2 + x^T \frac{\partial q(x)}{\partial x}} \left( \frac{\partial f}{\partial x} - A \right).$$

Matrix  $W(x)$  being positive definite under condition (33) can be concluded by applying lemmas 7 and 6.  $\diamond$ .



It follows from Lemma 6 that stability condition (22) can be guaranteed if  $r(t) = q(x)$  is chosen such that

$$e^{q(x)} \geq \lambda_{\max}(P_0). \quad (34)$$

It is obvious that both inequalities (33) and (34) can *always* be satisfied by simply choosing  $q(x)$  large enough and nonlinear enough. For square systems with  $\lambda_{\min}(BR_0^{-1}B^T) > 0$ , one can find  $q(x)$  without solving any algebraic Riccati equation since, by lemma 5, inequalities (33) and (34) can be rewritten as

$$2 + x^T \frac{\partial q(x)}{\partial x} > \frac{1}{2} \left\| \frac{\partial f}{\partial x} - A \right\| \cdot \frac{\lambda_{\max}(Q_0)}{\lambda_{\min}(Q_0)} \cdot \left[ \sqrt{\|A\|^2 + \lambda_{\min}(BR_0^{-1}B^T)\lambda_{\max}(Q_0)} - \|A\| \right],$$

and

$$e^{q(x)} \geq \frac{\lambda_{\max}(Q_0)}{\sqrt{\|A\|^2 + \lambda_{\min}(BR_0^{-1}B^T)\lambda_{\max}(Q_0)} - \|A\|}.$$

Based on lemma 8 and the discussion following (29), weighting matrices  $Q$  and  $R$  should be chosen as follows.

**Procedure for finding  $Q$  and  $R$ :** At any instant of time, consider the two complementary cases. If  $x \in \Omega^\circ$ , set  $x = 0$  in  $Q(x)$  and  $R(x)$  (or simply stop updating them). If  $x \notin \Omega^\circ$ , proceed with the following steps to find  $Q(x)$  and  $R(x)$ .

- Define  $Q(x)$  and  $R(x)$  according to (5) and with  $r(x) = q(x)$ .
- Find  $q(x)$  (on-line or off-line) such that inequalities (33) and (34) are met.

Summarizing our analysis in this and the preceding subsections, we can conclude the following theorem as one of the main results of this report.

**Theorem 1:** Assume that parameterizations  $A_i(x)$  be stabilizable for  $i = 1, \dots, L + 1$ . Then, using the proposed SDARE procedure (including the procedures of picking  $\alpha(x)$ ,  $Q(x)$  and  $R(x)$ ), the closed loop system under control (27) is globally asymptotically stable.

The second theorem establishes the fact that the proposed SDARE method generates near optimal controls. Although the theorem requires that matrix  $A(0, \alpha^\circ) - BR^{-1}(0)BP(0, \alpha^\circ)$  is diagonalizable, it is obvious that a slightly different upper bound can be developed for the case that the matrix is not diagonalizable.

**Theorem 2:** Assume that parameterizations  $A_i(x)$  be stabilizable for  $i = 1, \dots, L + 1$ . Then, using the proposed SDARE procedure (including the procedures of picking  $\alpha(x)$ ,  $Q(x)$  and  $R(x)$ ), the control (27) is near optimal in the sense that, if  $\Omega^\circ \subset \{x : \|x\| \leq d_\Omega\}$  for some small constant  $0 < d_\Omega \ll 1$  and if matrix  $A(0, \alpha^\circ) - BR^{-1}(0)BP(0, \alpha^\circ)$  is diagonalizable,

$$\begin{aligned} J^*(x(t_0), u^*, t_0, t_f) &\leq \frac{1}{2} x^T(t_0) P(x(t_0), t_0, \alpha^*(t_0)) x(t_0) \\ J(x(t_0), u, t_0, t_f) &\leq \frac{1}{2} x^T(t_0) P(x(t_0), t_0, \alpha^*(t_0)) x(t_0) + \frac{1}{2} x^T(t_f) \{P(x^*(t_f), t_f, \alpha^*(t_f)) \end{aligned}$$

$$\begin{aligned}
& -P(x(t_f), t_f, \alpha(t_f))\} x(t_f) \\
& + \frac{1}{2} x^T(t_0) \{P(x(t_0), t_0, \alpha(t_0)) - P(x(t_0), t_0, \alpha^*(t_0))\} x(t_0) \\
& + \frac{\lambda_{max}[Q(0) + P(0, \alpha^0)BR^{-1}(0)B^T P(0, \alpha^0)]}{4 \min_{1 \leq i \leq n} \{-\lambda_i[A(0, \alpha^0) - BR^{-1}(0)BP(0, \alpha^0)]\}} \cdot \|\Gamma\|^2 \cdot \|\Gamma^{-1}\|^2 \cdot d_\Omega^2, \\
& \qquad \qquad \qquad \forall t_f \geq t_0, \tag{35}
\end{aligned}$$

where  $\Gamma$  is the similarity transformation matrix that diagonalizes matrix  $A(0, \alpha^0) - BR^{-1}(0)BP(0, \alpha^0)$ .

Proof: Suppose that the system trajectory enters the set  $\Omega^0$  at time  $t = t_s$ . If  $t_f \in [t_0, t_s]$ , it follows from lemma 8 and from (4) that

$$J^*(x(t_0), u^*, t_0, t_f) \leq \frac{1}{2} x^T(t_0) P(x(t_0), t_0, \alpha^*(t_0)) x(t_0),$$

and that

$$S = P(x^*(t_f), t_f, \alpha^*(t_f)),$$

where  $\alpha^*(t)$  is the optimal solution to the two-point boundary-value problem, and  $x^*(t_f)$  is the final state of the optimal trajectory. During the same time interval, optimality condition is imposed by the choice of  $\alpha(\cdot)$ , therefore we have from (20) that

$$J(x(t_0), u, t_0, t_f) \leq \frac{1}{2} x^T(t_0) P(x(t_0), t_0, \alpha(t_0)) x(t_0) + \frac{1}{2} x^T(t_f) \{S - P(x(t_f), t_f, \alpha(t_f))\} x(t_f).$$

Obviously, the difference between upper bounds of  $J(\cdot)$  and  $J^*(\cdot)$  is due to the mismatch between one-point boundary-value solution and the optimal two-point boundary value solution.

If  $t_f > t_s$ , the difference between  $J(\cdot)$  and  $J^*(\cdot)$  will become larger in general than that during the time interval  $[t_0, t_s]$ . This is because the optimality condition is no longer imposed for  $t < t_s$ . Possible increase in the performance index for  $t > t_s$  can be calculated. It follows from (29) that, for  $t \geq t_s$ ,

$$\begin{aligned}
x &= e^{[A(0, \alpha^0) - BR^{-1}(0)BP(0, \alpha^0)]t} x(t_s) + e^{[A(0, \alpha^0) - BR^{-1}(0)BP(0, \alpha^0)]t} * O(\|x\|^2) \\
&\approx e^{[A(0, \alpha^0) - BR^{-1}(0)BP(0, \alpha^0)]t} x(t_s),
\end{aligned}$$

from which we have

$$\begin{aligned}
\|x\| &\approx \left\| e^{[A(0, \alpha^0) - BR^{-1}(0)BP(0, \alpha^0)]t} \right\| \cdot \|x(t_s)\| \\
&\leq \|\Gamma\| \cdot \|\Gamma^{-1}\| \cdot d_\Omega \cdot e^{-\min_{1 \leq i \leq n} \{-\lambda_i[A(0, \alpha^0) - BR^{-1}(0)BP(0, \alpha^0)]\}t}.
\end{aligned}$$

Applying this result in calculating performance index yields

$$\begin{aligned}
& \frac{1}{2} \int_{t_s}^{\infty} (x^T Q x + u^T R u) dt \\
&= \frac{1}{2} \int_{t_s}^{\infty} x^T [Q(0) + P(0, \alpha^0)BR^{-1}(0)BP(0, \alpha^0)] x dt
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{1}{2} \lambda_{\max} [Q(0) + P(0, \alpha^o) B R^{-1}(0) B^T P(0, \alpha^o)] \int_{t_*}^{\infty} \|x\|^2 dt \\
&\leq \frac{1}{2} \lambda_{\max} [Q(0) + P(0, \alpha^o) B R^{-1}(0) B^T P(0, \alpha^o)] \cdot \|\Gamma\|^2 \cdot \|\Gamma^{-1}\|^2 \cdot d_{\Omega}^2 \\
&\quad \cdot \int_{t_*}^{\infty} e^{-2 \min_{1 \leq i \leq n} \{-\lambda_i [A(0, \alpha^o) - B R^{-1}(0) B P(0, \alpha^o)]\} t} dt \\
&\leq \frac{\lambda_{\max} [Q(0) + P(0, \alpha^o) B R^{-1}(0) B^T P(0, \alpha^o)]}{4 \min_{1 \leq i \leq n} \{-\lambda_i [A(0, \alpha^o) - B R^{-1}(0) B P(0, \alpha^o)]\}} \cdot \|\Gamma\|^2 \cdot \|\Gamma^{-1}\|^2 \cdot d_{\Omega}^2 \\
&\quad \cdot e^{-2 \min_{1 \leq i \leq n} \{-\lambda_i [A(0, \alpha^o) - B R^{-1}(0) B P(0, \alpha^o)]\} t_*}.
\end{aligned}$$

The proof is completed by simply summarizing the above discussions.  $\diamond$

The upper bound of  $J$  can be explained as follows. The first two extra terms on the right hand side of (35) are due to the mismatch of terminal condition (since only one-point boundary-value problem is solved) and, if  $t_f = \infty$ , the first extra term will disappear (since  $x(\infty) = 0$ ). The third extra term on the right hand side of (35) exists because, in the set  $\Omega^o$ , the optimality condition is switched off and the resulting control is no longer optimal. Nevertheless, for regulation problem with  $t_f = \infty$ , the third extra term is small since design parameter  $d_{\Omega}$  is small and hence the difference between the optimal and sub-optimal performance indices is mainly the mismatch of terminal condition. This verifies the claim that the proposed control is near optimal.

### 3.3 Scalar Systems

To gain some insight about the choice of Lyapunov function and the choices of matrices  $Q$  and  $R$ , let us study their implications to scalar systems. In addition, scalar systems have several interesting features that are distinct from multivariable systems.

Consider the scalar system:

$$\dot{x} = a(x)x + b(x)u,$$

where  $b(x) \neq 0$ . As discussed before, we can assume that  $b(x) = 1$  since it can always be achieved under the state transformation

$$z = \int \frac{1}{b(x)} dx.$$

Therefore, we need only to study the system

$$\dot{x} = a(x)x + u. \quad (36)$$

For system (36), the state-dependent algebraic Riccati equation is, for any  $q \triangleq Q > 0$  and  $r \triangleq R > 0$ ,

$$2ap - \frac{1}{r}p^2 + q = 0,$$

and the optimal control candidate is

$$u = -\frac{1}{r}px.$$

The positive solution of the algebraic Riccati equation is

$$p(x) = r \left[ a + \sqrt{a^2 + \frac{q}{r}} \right].$$

Substituting this solution and the control into system (36) yields the closed loop optimal system

$$\dot{x} = -\sqrt{a^2 + \frac{q}{r}}x.$$

It follows from Lyapunov function

$$V = \int xp(x)dx$$

that

$$\dot{V} = xp(x)\dot{x} = -r \left[ a + \sqrt{a^2 + \frac{q}{r}} \right] \sqrt{a^2 + \frac{q}{r}} x^2$$

is negative definite and therefore the system is globally asymptotically stable. It follows from solution  $p$  that

$$\begin{aligned} \dot{p} &= r \left[ 1 + \frac{a}{\sqrt{a^2 + \frac{q}{r}}} \right] \dot{a} + \frac{1}{2\sqrt{a^2 + \frac{q}{r}}} \dot{q} + \left( a + \frac{2a^2r + q}{2\sqrt{a^2r^2 + qr}} \right) \dot{r} \\ &= -r \left[ \sqrt{a^2 + \frac{q}{r}} + a \right] x \frac{da}{dx} - \frac{1}{2} x \frac{dq}{dx} - \frac{1}{2} \left[ a + \sqrt{a^2 + \frac{q}{r}} \right]^2 x \frac{dr}{dx} \\ &= -px \frac{da}{dx} - \frac{1}{2} x \frac{dq}{dx} - \frac{p^2}{2r^2} x \frac{dr}{dx}, \end{aligned}$$

which, by the algebraic Riccati equation, is the optimality condition (16). In summary, we have the following result:

**Lemma 9:** For scalar systems, the SDARE method always yields the optimal control and the optimal control is globally stabilizing.

Essentially, lemma 9 holds because scalar systems in the form (36) are always stabilizable and all its dynamics are matched (that is, in the span of the input matrix). Mathematically, stability proof is straightforward because it couples the algebraic Riccati equation with the Lyapunov function whose time derivative is  $\dot{V} = xp\dot{x}$ . For multivariable systems, global stability could be concluded in the similar fashion without any condition imposed if we would find a Lyapunov function  $V$  such that  $\dot{V} = x^T P \dot{x}$  where  $P$  is the solution to the algebraic Riccati equation. This means that  $V$  should have the property that

$$\frac{\partial V}{\partial x} = x^T P \quad \Rightarrow \quad \frac{\partial V}{\partial x_i} = x^T P_i,$$

where  $P_i$  is the  $i$ -th row of  $P$ . Since  $V$  is scalar function, second order partial derivatives of  $V$  must commute, that is, for all  $i, j \in \{1, \dots, n\}$ ,

$$\frac{\partial^2 V}{\partial x_i \partial x_j} = \frac{\partial^2 V}{\partial x_j \partial x_i}.$$

The above two equations imply that the solution to the algebraic Riccati equation must have the property that, for all  $x \in \mathbb{R}^n$  and for all  $i, j \in \{1, \dots, n\}$ ,

$$x^T \frac{\partial P_i}{\partial x_j} = x^T \frac{\partial P_j}{\partial x_i}.$$

where  $P_i$  and  $P_j$  are the  $i$ -th and  $j$ -th rows of  $P$ . Obviously, the above condition is too restrictive to be satisfied for nonlinear systems. Therefore, one can not and should not expect that stability of general, multivariable, nonlinear optimal control systems can be concluded without any condition. In this sense, Lyapunov function (21) is the best due to its simplicity.

When Lyapunov function (21) is used, condition (24) has to be satisfied to conclude stability. Although this condition is conservative for scalar systems, it is needed for multivariable systems. Let us verify our way of choosing  $Q$  and  $R$  in section 3 by investigating scalar systems. There are two complementary cases.

Case 1: Both  $r(x)$  and  $q(x)$  are chosen to be constant.

It follows that

$$w = q + r \left[ a + \sqrt{a^2 + \frac{q}{r}} \right]^2 + r \left[ \sqrt{a^2 + \frac{q}{r}} + a \right] x \frac{da}{dx}.$$

For bounded dynamics  $a(x)$  that is not linear (for example,  $a(x) = e^{-x}$ ), the last term in  $w$  will grow unbounded while the first two terms are bounded, and hence  $w > 0$  can not be guaranteed globally (but only semi-globally). Therefore, for global stability, making both  $Q$  and  $R$  constant is not a good choice even for scalar systems.

Case 2: At least one of  $q(x)$  and  $r(x)$  is not constant.

It follows from the optimality condition that

$$\begin{aligned} w &= q + \frac{p^2}{r} + \frac{1}{2}x \frac{dq}{dx} + \frac{p^2}{2r^2}x \frac{dr}{dx} + px \frac{da}{dx} \\ &= q + \frac{1}{2}x \frac{dq}{dx} + \frac{p^2}{2r^2}x \frac{dr}{dx} + \left( \frac{p}{\sqrt{r}} + \frac{1}{2}\sqrt{r}x \frac{da}{dx} \right)^2 - \frac{1}{4}rx^2 \left( \frac{da}{dx} \right)^2, \end{aligned}$$

which can be made positive definite by simply choosing  $q(x) > 0$  such that

$$\frac{1}{4}rx^2 \left( \frac{da}{dx} \right)^2 < \frac{1}{2}x \frac{dq}{dx} + \frac{p^2}{2r^2}x \frac{dr}{dx} + q.$$

The existence of solutions to the above differential inequality is obvious.

In summary, the above discussion shows that the choices of Lyapunov function and weighting matrices in performance index used in the previous sections are appropriate. In other words, although the resulting control is sub-optimal, the proposed method is an excellent (if not the best) trade-off among optimality and efficiency in implementation by maintaining global stability and good performance.

## References

- [1] M.Athens and P.L.Falb, *Optimal Control*, McGraw Hill, New York, NY, 1966.
- [2] A.E.Bryson and Y.-C. Ho, *Applied Optimal Control*, 2nd edition, Hemisphere Publishing Corporation, New York, NY, 1975.

**PROCESS MODELING AND EXPERIMENTAL STUDIES OF  
THE DENSIFICATION OF POWDER COMPACTS**

**M. N. Rahaman  
Professor  
Department of Ceramic Engineering**

**University of Missouri-Rolla  
222 McNutt Hall  
Rolla, MO 65401-0249**

**Final Report for:  
Summer Faculty Research Program  
Wright Laboratory**

**Sponsored by:  
Air Force Office of Scientific Research  
Bolling Air Force Base, DC**

**and  
Wright Laboratory**

**August 1995**

# PROCESS MODELING AND EXPERIMENTAL STUDIES OF THE DENSIFICATION OF POWDER COMPACTS

M. N. Rahaman  
Professor  
Department of Ceramic Engineering  
University of Missouri-Rolla

## Abstract

Mechanistic and continuum models for the densification of powder compacts under a complex stress system were reviewed. The key features of the models were examined and their predictions for the effect of porosity on the densification rate were compared. Experimental studies of the sintering of mullite core particles coated with a layer of amorphous mullite were performed. The coated particles (in the form of a concentrated suspension) can potentially be used in the infiltration of fiber preforms for the production of porous matrix composites. Compared to uncoated mullite particles, the higher sinterability of the coating is expected to lead to a reduction in the fabrication temperature of the composite.

# PROCESS MODELING AND EXPERIMENTAL STUDIES OF THE DENSIFICATION OF POWDER COMPACTS

M. N. Rahaman

## Introduction

Investigations were performed in two areas: (i) process modeling of the densification of powder compacts and (ii) sintering of mullite core particles coated with an amorphous layer of mullite. These investigations are described separately in the following sections.

## 1. Process Modeling of the Densification of Powder Compacts

### 1.1 Background

Theoretical models for the densification of powder compacts are important technologically and scientifically. Ideally, the models can be used in the design of a suitable process cycle for the fabrication of a component without costly trial and error experiments. The models can also provide valuable information in the near net shape fabrication of a component that is subjected to a fairly complex stress system during densification (e.g., hot isostatic pressing of a powder encapsulated within a metal can). Furthermore, the predictions of the models, particularly when compared with the results of experiments performed under controlled conditions, can provide physical insight into the key mechanisms and parameters in the fabrication process.

Models for the densification of a powder compact subjected to a complex stress state predict an equation of the form:

$$\dot{\epsilon}_{ij} = \frac{\dot{\epsilon}_0}{\sigma_0} \left( \frac{G_0}{G} \right)^b \left\{ \frac{3}{2} c(\rho) s_{ij} + \frac{1}{3} f(\rho) [\sigma_m - \sigma_s] \delta_{ij} \right\} \quad (1)$$

where  $\dot{\epsilon}_{ij}$  is the inelastic strain rate produced by the applied stress  $\sigma_{ij}$ ,  $\delta_{ij}$  is the Kronecker delta,  $\sigma_m$  is the mean stress  $[= (1/3)\sigma_{kk}]$ ,  $\sigma_s$  is the sintering stress,  $s_{ij}$  is the stress deviator  $(= \sigma_{ij} - \sigma_m \delta_{ij})$ ,  $G$  is the grain size of the porous solid,  $\dot{\epsilon}_0$  is the strain rate of the fully dense solid with a grain size  $G_0$  subjected to a uniaxial stress  $\sigma_0$ ,  $b$  is a grain growth exponent and  $c(\rho)$  and  $f(\rho)$



are functions of the relative density. For example, under a pure hydrostatic stress,  $\sigma_{11} = \sigma_{22} = \sigma_{33} = \sigma$ . Assuming that  $\sigma_s \ll \sigma_m$  (valid for most experiments in hot isostatic pressing) and that the compact is isotropic, the densification rate is given by:

$$\frac{\dot{\rho}}{\rho} = - \dot{\epsilon}_{kk} = - \frac{\dot{\epsilon}_0}{\sigma_0} \left( \frac{G_0}{G} \right)^b f(\rho) \sigma \quad (2)$$

The models differ in the way  $c(\rho)$  and  $f(\rho)$  are determined and can be classified into two main types: (i) mechanistic models (also referred to as micromechanical models) and (ii) empirical models.

In the mechanistic models [1-10], a rough approximation to the microstructure of the powder compact is assumed and the equations for  $c(\rho)$  and  $f(\rho)$  are derived analytically for a specified mechanism of matter transport. Since the microstructure of the powder compact changes drastically during densification, a single geometrical model is inadequate to provide a reasonable approximation for the entire process. The densification process is therefore divided into two or three stages and an appropriate model for each stage is assumed. An advantage of the mechanistic models is that the derivations of the equations for  $c(\rho)$  and  $f(\rho)$  are fairly rigorous. However, a major problem is that the assumed geometrical model represents a drastic simplification of the microstructure of real powder compacts.

In the empirical models, the material is assumed to obey a constitutive equation and the functions  $c(\rho)$  and  $f(\rho)$  are found by curve fitting [11-14]. Serious errors can occur when the samples used for the determination of  $c(\rho)$  and  $f(\rho)$  differ in microstructural characteristics (e.g., grain size and pore size size distribution).

Recently, Dutton et al [15] used a different approach to determine the density dependent functions in Eq. (1). They assumed a yield function put forward by Doraivelu et al [16] for the plastic deformation of metals . Applying the principle of conservation of energy and incorporating a term to account for grain growth, Dutton et al derived the following equation:

$$\dot{\epsilon}_{ij} = \frac{\dot{\epsilon}_0}{\sigma_0} \left( \frac{G_0}{G} \right)^b \phi^2 \rho \left\{ (1 + \nu) s_{ij} + (1 - 2\nu) [\sigma_m - \sigma_s] \delta_{ij} \right\} \quad (3)$$

where  $\phi$  is a geometrical term that accounts for the change in cross-sectional area of the sample due to porosity and  $\nu$  is the Poisson's ratio. Application of a force balance across any cross-section gives:

$$\phi = A_0/A_e \quad (4)$$

where  $A_0$  is the total cross-sectional area (solid plus porosity) and  $A_e$  is the effective cross-sectional area in the solid phase.

Comparing Eqs. (1) and (3), it is seen that one result of the approach of Dutton et al is that the density-dependent terms  $c(\rho)$  and  $f(\rho)$  are separated into two parts: (i) a geometrical term,  $\phi^2$ , and (ii) a term that depends on the Poisson's ratio,  $\nu$ . An advantage of this approach is that both  $\phi$  and  $\nu$  can be measured simultaneously and relatively easily in uniaxial loading experiments on porous samples.

## 1.2 Expressions for the Density-Dependent Functions

A variety of expressions have been put forward for  $f(\rho)$  but corresponding expressions for  $c(\rho)$  are fewer. Taking a range of  $\rho$  between  $\approx 0.64$  to  $\approx 0.90$  (i.e., an initial density corresponding to that of dense random packing of monosize, spherical particles and an end-point density when the pores become isolated) where the effect of porosity is known to have a significant effect on the strain rate, expressions for  $f(\rho)$  are summarized below. For this density range, a good approximation for  $c(\rho)$  is [7]:

$$\frac{f(\rho)}{c(\rho)} = 0.5 \quad (5)$$

(a) Helle et al [3]: Mechanistic model consisting of spherical, monosize particles in dense random packing where matter transport occurs by diffusion:

$$f(\rho) = \frac{0.5 (1 - \rho_0)^2}{\rho (\rho - \rho_0)^2} \quad (6a)$$

where  $\rho_0$  is the initial relative density ( $\approx 0.64$ )

(b) Beeré [17]: Mechanistic model consisting of a truncated octahedron (tetraikaidecahedron) with pores along the grain edges and involving a minimization of the total interfacial area:

$$f(\rho) = \exp[a(1 - \rho)] \quad (6b)$$

where  $a$  is a constant that depends on the equilibrium dihedral angle. For a dihedral angle of  $120^\circ$  (equal to the measured value for high-purity alumina),  $a \approx 4$ .

(c) Scherer [18]: Mechanistic model consisting of cylinders in cubic arrangement where matter transport occurs by viscous flow:

$$f(\rho) = \frac{(3 - 2\rho)}{\rho} \left[ 1 - \left( \frac{\rho}{3 - 2\rho} \right)^{1/2} \right] \quad (6c)$$

(d) Du and Cocks [7]: Empirical analysis of the torsional creep data of Coble and Kingery [19] for alumina and making use of Eq. [5]:

$$f(\rho) = 0.5 \rho^{-5.3} \quad (6d)$$

(e) Shima and Oyane [13]: Empirical analysis of the stress-strain data for powder compacts of copper:

$$f(\rho) = \frac{2.5 (1 - \rho)^{0.5}}{\rho^{2.5}} \quad (6e)$$

The values for  $f(\rho)$  determined from Eqs. (6a) to (6e) are plotted in Fig. 1. The empirical equation of Du and Cocks and the analysis by Beeré give roughly the same values for  $f(\rho)$  for alumina. These  $f(\rho)$  values for alumina obtained from the model of Beeré and from the empirical analysis of Du and Cocks are very close to those for copper determined empirically by Shima and Oyane using a curve-fitting technique. The assumption of the Shima and Oyane expression by Dutton et al [15] in modeling the densification behavior of alumina appears reasonable.

The  $f(\rho)$  values for glass determined from the model of Scherer are significantly lower than those for the other models. Finally, the mechanistic model of Helle et al gives  $f(\rho)$  values that are significantly higher than those for the other models. The use of the expression of Helle et al should therefore lead to predictions of deformation rates that are significantly enhanced compared to those obtained with Eqs. (6b)-(6e).

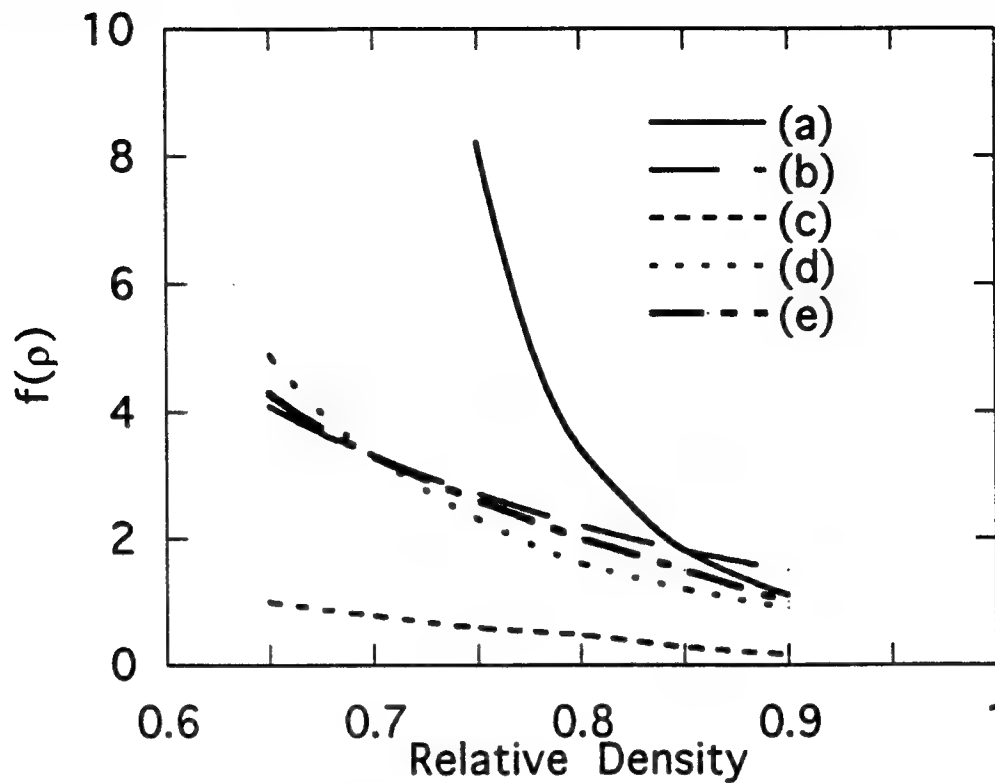


Figure 1. Values for the density-dependent function  $f(\rho)$  determined from Eqs. 6(a) to 6(e).

## 2. Sintering of Coated Mullite Particles

### 2.1 Background

Considerable difficulties are commonly encountered in the fabrication of fiber reinforced ceramic matrix composites with the desired properties and reliability for mechanical engineering applications at elevated temperatures. The desired microstructure of the composite is ideally a fully dense matrix containing a uniform arrangement of fibers. However, the production of such a microstructure is commonly expensive or may require fabrication temperatures that lead to degradation of the fiber and interfacial properties. Furthermore, the large shrinkage that must normally accompany the densification of the matrix can impose significant stresses on the fibers, thereby leading to microstructural damage. Porous matrix composites offer a trade-off between the properties of the composite and the ease of fabrication. The properties of the porous

composite are not as good as those of a similar system with a dense matrix but the fabrication process can present fewer difficulties. When the reinforcing fibers have been woven to produce a three-dimensional preform, a common fabrication route involves incorporation of the matrix by infiltration of reacting gases (chemical vapor infiltration), of a molten pre-ceramic polymer, or of a powder slurry.

The composite system and fabrication route of interest to the present work involves infiltration of a Nextel fiber preform with a slurry of mullite particles. After the infiltration process, the matrix phase consists essentially of particles in point contact and therefore possesses very little strength. The particles must be bonded together at the contact points (i.e., "necks" must be produced between touching particles) to provide a reasonable amount of matrix strength and to preserve the overall integrity of the composite. Another requirement is that the matrix must not undergo significant shrinkage in order to prevent the imposition of large stresses on the fibers. A further requirement is that the temperature required to produce adequate necking between the particles must be below the temperature at which the strength of the fibers start to degrade. In the present system, the fabrication temperature ( $\approx 1300^\circ\text{C}$ ) required to achieve the desired microstructural change in a reasonable time is higher than that ( $\approx 1100^\circ\text{C}$ ) at which the strength of the fiber starts to degrade.

One technique for reducing the fabrication temperature of the composite involves coating the mullite core particles with a layer of highly sinterable material in which matter transport can be accomplished more readily than that in the core mullite particles. The use of coated particles for improving the sintering characteristics of particulate systems is now an established technique in ceramic fabrication [20-23]. In the present system, the mullite core particles were coated with an amorphous mullite layer by precipitation from a mullite sol. (Because of the lower viscosity, amorphous materials have a higher sinterability than that of crystalline materials of the same chemical composition.) The thickness of the coating must, however, be small (i.e., much less

than the radius of the core particles) so that the overall shrinkage of the matrix phase of the composite can be kept small.

The objective of the present investigations was to provide an understanding of the sintering characteristics of mullite core particles coated with an amorphous mullite layer in order to determine the feasibility of using the coated particles for achieving the fabrication requirements of the Nextel fiber/mullite matrix composite. The overall approach can be divided into two parts. In the first part, core particles with a thick enough coating would be used to demonstrate the effectiveness of the coating technique and to measure the effect of the coating on the sintering characteristics of the core particles. In the second part, the understanding gained in the first part would be used to optimize the coating thickness and the sintering schedule in order to meet the fabrication requirements and to achieve adequate strength in the fabricated porous compact.

## 2.2 Experimental Procedure

Powder compacts ( $\approx 5$  mm in diameter by 5 mm) of the coated mullite particles, the uncoated mullite particles (average particle size  $\approx 0.2$   $\mu\text{m}$ ) and the amorphous mullite coating were formed by die pressing followed by cold isostatic pressing (40,000 psi). For the coated mullite particles, complete drying of the powder prior to compaction resulted in the formation of relatively hard agglomerates. Hard agglomerates are difficult to compact and also lead to considerable heterogeneity in the packing of the powder compact. The coated particles were therefore die-pressed in a semi-dry state and dried overnight in an oven ( $\approx 100$   $^{\circ}\text{C}$ ) prior to cold isostatic pressing. The uncoated powder and the coating material were die-pressed in the dry state.

In the initial sintering experiments to observe the overall sintering characteristics of the powders, the compacts were sintered at  $5^{\circ}\text{C}/\text{min}$  to  $1500$   $^{\circ}\text{C}$ . Sintering was carried out in air in a dilatometer that allowed continuous monitoring of the shrinkage kinetics. Powder compacts of the coated particles and the uncoated particles were also sintered for 24 h at  $1100$   $^{\circ}\text{C}$  to

investigate the potential of lower temperature annealing for enhancing the strength of the compacts.

The microstructures of polished surfaces of the sintered compacts were observed in the scanning electron microscope (SEM). Because of their highly porous nature, the compacts were vacuum impregnated with epoxy resin prior to polishing.

### 2.3 Results and Discussion

Figure 2 shows the data for linear shrinkage,  $\Delta L/L_0$ , versus temperature for the powder compacts formed from the uncoated mullite particles and the coated mullite particles during constant heating rate sintering in air at 5 °C/min to 1500 °C. ( $L_0$  = initial sample length and  $\Delta L = L_0 - L$ , where  $L$  is the instantaneous sample length.) Apart from a small difference at  $\approx 300$ -500 °C which is most likely due to a difference in weight loss, the curves are almost identical. The sintering results therefore indicate that the coated powders have only a very thin layer of coating or are not properly coated. Transmission electron microscopy is currently being used to observe the nature of the coating on the core particles.

Figure 2 also shows that the shrinkage increases rapidly above  $\approx 1300$  °C and reaches a value of 4-5% at 1500 °C. However, the compacts suffer a weight loss of  $\approx 10\%$  during sintering. Assuming that the weight loss,  $\Delta m/m_0$ , gives rise to an equivalent volume change,  $\Delta V/V_0$ , according to the relation:

$$\frac{\Delta m}{m_0} = \frac{\Delta V}{V_0} = 3 \frac{\Delta L}{L_0} \quad (7)$$

then a significant part of the observed linear shrinkage ( $\approx 3\%$ ) can be accounted for in terms of the weight loss. The final densities of the sintered compacts are therefore only slightly higher than the initial (i.e., green) densities ( $\approx 1.7 \text{ g/cm}^3$  or  $\approx 0.53$  of the theoretical density of mullite).

As outlined earlier, in separate experiments, the powder compacts were sintered for 24 h at 1100 °C. Preliminary measurements indicate that the compacts formed from the coated

powders are stronger than the compacts formed from the uncoated powders but not significantly stronger than the isostatically-pressed green compacts.

SEM of the polished surfaces of the compacts formed from the coated powder and the uncoated powder revealed no significant difference in the microstructure after sintering to 1500 °C. However, the microstructure was heterogeneous, with large density variations from one region to another. Figure 3 shows a region of fairly good packing in the sintered compact formed from the coated particles. The average distance between the pores (interpore distance) is much larger than the initial size ( $\approx 0.2 \mu\text{m}$ ) of the mullite core particles. This may indicate that the coated particles consist of agglomerates (rather than single particles). The agglomerates undergo local densification leading to an interpore distance characteristic of the size of the agglomerates.

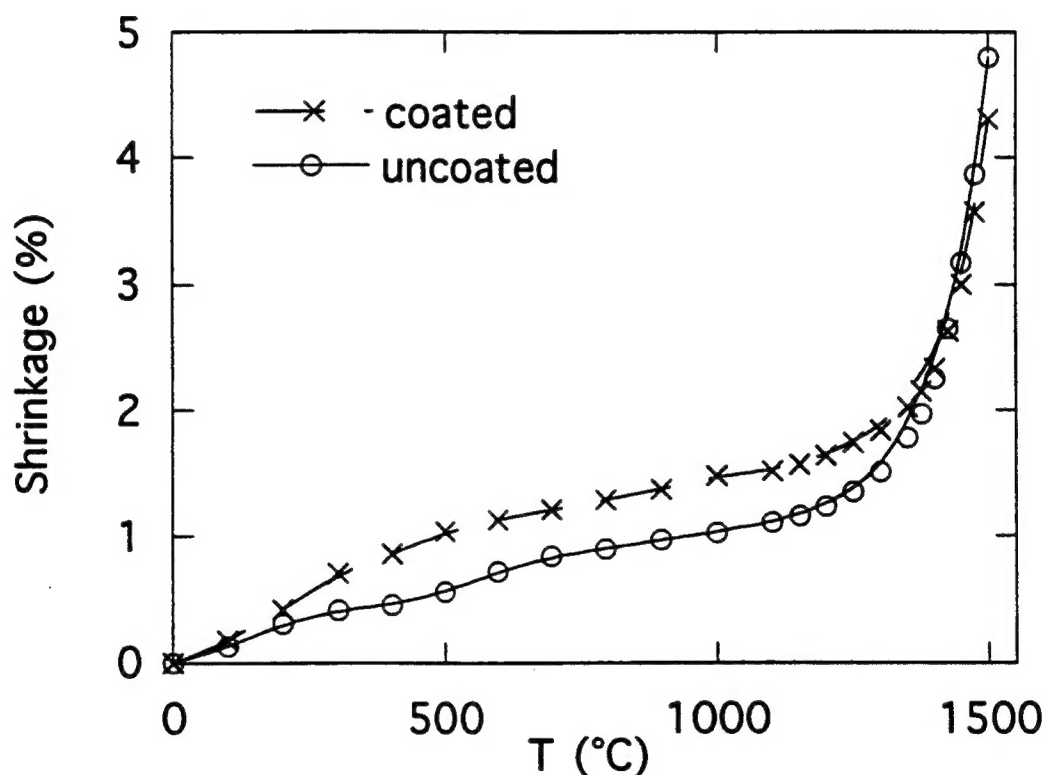


Figure 2. Shrinkage versus temperature for the powder compacts formed from the coated mullite powders and the uncoated mullite powders during constant heating rate sintering at 5°C/min to 1500 °C.



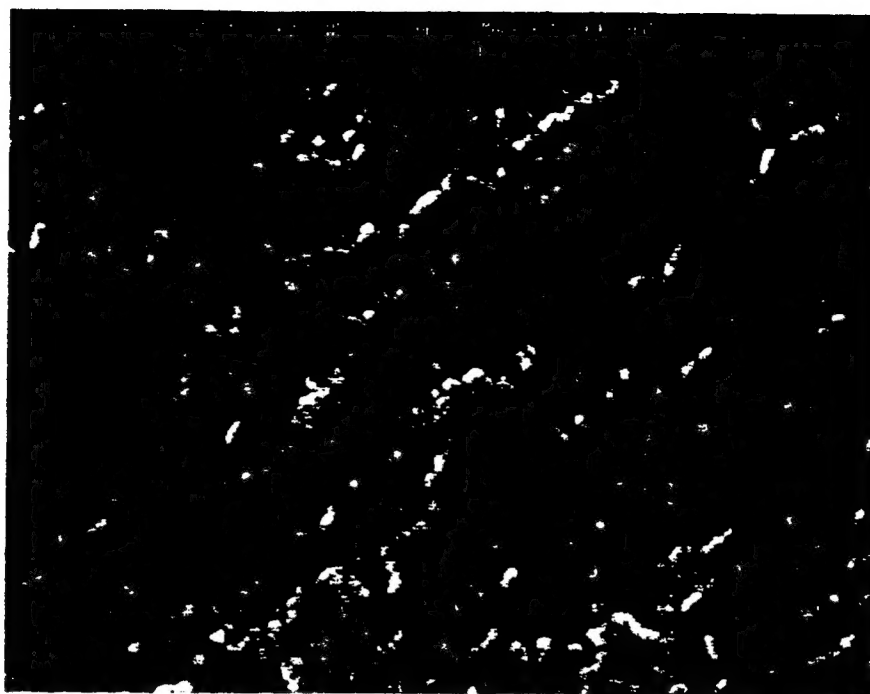


Figure 3. Scanning electron micrograph of the polished surface of a powder compact formed from the coated mullite powder after sintering at 5°C/min to 1500 °C.

The preliminary experiments performed so far indicate the need for better control of the coating process for the preparation of the coated mullite particles. Further experiments are being performed to optimize the coating process and to investigate the effects of the coating on the sintering and strength of the compacts.

### Conclusions

Mechanistic and empirical models for the densification of powder compacts give widely different predictions for the effect of porosity on the inelastic strain rate. The results of theoretical simulations of the densification process are expected to be controlled not only by the assumed values of the physical constants (e.g., diffusion coefficients) but also by the assumed density-

dependent functions. Reliable data for the physical constants and the density-dependent functions are needed to test the usefulness of the models for predicting the densification and deformation of practical systems.

Experiments on the sintering of mullite core particles coated with a layer of amorphous mullite show no significant differences between the compacts formed from the coated powders and those formed from the core particles. The data indicate that the coated particles have a very thin layer of coating or are not properly coated. Optimization of the coating process and the characterization of the coated powders are needed to successfully exploit the technique for the present system.

### References

1. M. F. Ashby, *Acta metall.*, 22 275 (1974).
2. F. B. Swinkels and M. F. Ashby, *Acta metall.*, 29 259 (1981).
3. A. S. Helle, K. E. Easterling, and M. F. Ashby, *Acta metall.*, 33 2163 (1985).
4. E. Arzt, M. F. Ashby, and K. E. Easterling, *Metall. Trans. A*, 14A 211 (1983).
5. R. M. McMeeking and L. T. Kuhn, *Acta metall. mater.*, 40 961 (1992).
6. A. C. F. Cocks, *Acta metall. mater.*, 42 2191 (1994).
7. Z.-Z. Du and A. C. F. Cocks, *Acta metall. mater.*, 40 1969 (1992).
8. Z.-Z. Du and A. C. F. Cocks, *Acta metall. mater.*, 40 1981 (1992).
9. Y.-M. Liu, H. N. G. Hadley, and J. M. Duva, *Acta metall. mater.*, 42 2247 (1994).
10. D. M. Elzey and H. N. G. Hadley, *Acta metall. mater.*, 41 2297 (1993).
11. J. Besson and M. Abouaf, *J. Am. Ceram. Soc.*, 75 2165 (1992).
12. J. Besson, F. Valin, P. Lointier, and M. Boncoeur, *J. Mater. Eng. Perf.*, 1 637 (1992).
13. S. Shima and M. Oyane, *Int. J. Mech. Sci.*, 18 285 (1976).
14. D. N. Lee and H. S. Kim, *Powder Metall.*, 35 275 (1992).
15. R. E. Dutton, S. Shamasundar, and S. L. Semiatin, *Metall. Trans. A*, August, 1995.

16. S. M. Doraivelu, H. L. Gegel, J. S. Gunasekera, J. C. Malas, J. T. Morgan, and J. F. Thomas, Jr., *Int. J. Mech. Sci.*, 26 527 (1984).
17. W. Beeré, *Acta metall.*, 23 131 (1975).
18. G. W. Scherer, *J. Am. Ceram. Soc.*, 60 239 (1977).
19. R. L. Coble and W. D. Kingery, *J. Am. Ceram. Soc.*, 39 377 (1956).
20. D. Kapolnek and L. C. De Jonghe, *J. Europ. Ceram. Soc.*, 7 345 (1991).
21. M. D. Sacks, N. Bozkurt, and G. W. Scheiffele, *J. Am. Ceram. Soc.*, 74 2428 (1991).
22. C.-L. Hu and M. N. Rahaman, *J. Am. Ceram. Soc.*, 75 2066 (1992).
23. C.-L. Hu and M. N. Rahaman, *J. Am. Ceram. Soc.*, 76 2549 (1993).